Model complexity and out-of-sample performance: evidence from S&P 500 index returns

Article (Accepted Version)

Model Complexity and Out-of-Sample Performance: Evidence from S&P 500 Index Returns

Andreas Kaeck[§] **Paulo Rodrigues**[‡] **Norman J. Seeger**[†]

[a] *University of Sussex, United Kingdom, E-mail: ak486@sussex.ac.uk*
[b] *Maastricht University, The Netherlands, E-mail: p.rodrigues@maastrichtuniversity.nl*
[c] *Corresponding author: VU University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands, E-mail: n.j.seeger@vu.nl, Phone: +31 20 598 1512*

[§]University of Sussex, United Kingdom, E-mail: ak486@sussex.ac.uk,
Phone: +44 1273 678433
[‡]Maastricht University, The Netherlands, E-mail: p.rodrigues@maastrichtuniversity.nl,
Phone: +31 43 388 36 33
[†]Corresponding author. VU University Amsterdam, The Netherlands, E-mail: n.j.seeger@vu.nl,
Phone: +31 20 59 81512

Abstract. We apply a range of *out-of-sample* specification tests to more than forty competing stochastic volatility models to address how model complexity affects out-of-sample performance. Using daily S&P 500 index returns, model confidence set estimations provide strong evidence that the most important model feature is the non-affinity of the variance process. Despite testing alternative specifications during the turbulent market regime of the global financial crisis of 2008, we find no evidence that either finite- or infinite-activity jump models or other previously proposed model extensions improve the out-of-sample performance further. Applications to Value-at-Risk demonstrate the economic significance of our results. Furthermore, the out-of-sample results suggest that standard jump diffusion models are misspecified.

Key Words: Out-of-sample specification tests; jump-diffusion models; Lévy-jump models; non-affine variance models; forecasting

JEL Classifications: G12; G15; C53

1 Introduction

In this paper, we analyze continuous-time and discrete-time models for S&P 500 index returns to study the relationship between model complexity and out-of-sample performance. The study of time-series dynamics of major stock market indices, such as the S&P 500, has previously attracted a large number of empirical studies, see e.g. Eraker *et al.* (2003), Christoffersen *et al.* (2010), Bates (2012), or Kou *et al.* (2013) and applied research today is faced with the challenge of selecting model dynamics from a huge number of alternative specifications.

Despite the importance of this research area, many papers in the continuous-time literature focus on *in-sample* specification tests. In this paper, we diverge from this approach and provide a range of different *out-of-sample* performance tests. In-sample studies are very helpful to learn about the structural building blocks required to produce stylized facts in the data. However, eventually the out-of-sample performance of a model is crucial for market participants using such a model in finance applications that are affected by uncertain future market scenarios. Our main aim is to understand to what extent the superior performance of sophisticated stochastic models prevails when they are applied outside their estimation period. To this end, we first estimate more than forty different stochastic models that encompass the most widely used model features in the continuous-time literature. These features include affine vs non-affine models, single-factor vs multi-factor specifications, diffusion models vs jump models, finite activity vs infinite activity, discrete-time vs continuous-time models. The combination of these building blocks leads to a very comprehensive set of competing models. Although not the focus of this paper, we also study some new model specifications such as non-affine time-changed Lévy models. To the best of our knowledge, this paper is the first to provide comprehensive out-of-sample evidence for such a large set of stochastic models.[1]

[1]Few papers consider the out-of-sample performance of continuous-time models. Yun (2014) conducts a range of density forecasting tests using affine one-factor jump-diffusion models, Shackleton *et al.* (2010) use similar model specifications. This paper differs substantially from the aforementioned papers as we

Various model specification tests are then applied to an out-of-sample period of S&P 500 index returns, including the turbulent market regime during the onset of the financial market crisis in 2008. To compare performances of a very large number of model specifications, we employ the model confidence set estimation procedure of Hansen *et al.* (2011). We separate subsets of models that have a statistically indistinguishable performance according to various different out-of-sample loss functions. In doing so, we accept that one single best performing model might not exist but rather that different modeling approaches may be equally successful. First, we compare likelihood-based out-of-sample fit statistics, including sequential likelihoods as proposed by Johannes *et al.* (2009). This allows us to detect the time-periods during which particular model specifications out- or underperform. Secondly, we follow Gneiting and Ranjan (2011) in comparing models using the continuous ranked probability score (CRPS), a criterion that can be used to compare the out-of-sample forecasting performance. CRPS has the advantage that weighted versions of the statistic retain propriety, which is essential for comparing the performance in various areas of the forecasting distributions. It is often argued that jump models in particular provide a better fit to the tails of the return distribution, and weighted CRPS fit statistics are employed to study model performance in the tails (as well as the center of the return distribution). Thirdly, we test the economic significance of our results by applying the VaR loss function of González-Rivera *et al.* (2004). And fourthly, we use a range of absolute model tests suggested by Berkowitz (2001) and others.

Our empirical tests provide two main results. First, we find that no model is able to produce out-of-sample predictions in line with the true data-generating process. Using the test statistics developed in Berkowitz (2001) we find that all models analyzed are rejected when tested on the entire out-of-sample period. Second, we find that in terms of relative model performance more parsimonious stochastic volatility models outperform models that include a jump component. This is a surprising result, since numerous papers find that jump models outperform continuous stochastic volatility models *in-sample* (see

focus on a much broader number of models and specification tests.

Eraker *et al.* (2003), Eraker (2004) or Ignatieva *et al.* (2015)).

There are two possible explanations why jump models are outperformed. First, one may interpret this result as evidence for misspecification of the jump component (despite the fact that we use quite sophisticated jump modeling) and *not* as evidence against the importance of modeling jumps in equity returns. Our results may be driven by the fact that jumps are difficult to estimate and jump distributions and intensities may vary strongly over time. For instance, jump parameters may be very different during periods of crisis and this may cause model misspecification. This finding is related to results in Santa-Clara and Yan (2010) who find a weak connection between variance and the jump intensity when both processes are estimated independently.

Second, and more important, our results may provide useful insights of how the global financial crisis unfolded. High returns may either be driven by jumps or stochastic volatility. Jumps are crucial to explain a number of rare events such as the market crash of 1987 (see the discussion in Eraker *et al.* (2003)). On the contrary, periods of high market volatility may render jumps obsolete as stochastic volatility is sufficient to generate a sequence of large returns in times of prolonged high market volatility. The result of pure stochastic volatility models outperforming jump models implies that an increasing level of market volatility during our out-of-sample period was sufficient to model financial crisis returns from 2007 to 2009. The distinction between how shocks are created is important for many applications in finance as rare event models may have very different implications compared to models driven by stochastic volatility. This finding is related to Stroud and Johannes (2014) who draw similar conclusions using high frequency returns.

Finally, we provide several additional empirical exercises to corroborate our findings. First, we increase our parameter updating frequency to investigate the impact of the partitioning of the sample into one estimation and one forecasting period. Second, we investigate the effect of including additional data in the information set, namely realized variance and the VIX index. Third, we investigate the impact of time varying expected returns on our results.

2 Related Literature

Prior literature on testing continuous time models for stock returns are often interested in the in-sample performance of models. To tackle the challenge of estimating complex continuous time models a range of different estimation and filtering techniques has been developed. These include simulated methods of moments approaches, approximate maximum likelihood estimation, efficient methods of moments and Bayesian MCMC estimation algorithms (see Andersen *et al.*, 2002, Eraker *et al.*, 2003, Bates, 2006 or Johannes *et al.*, 2009). At least partly driven by the differences in estimation methodology, there appears to be no standard in the continuous-time literature as far as model evaluation criteria are concerned. Eraker *et al.* (2003) use Bayes Factors and in-sample QQ plots to assess the in-sample fit and provide evidence of the impact of several model features on the shape of implied volatility smiles. Bates (2012) also uses in-sample QQ plots and a comparison of in-sample unconditional distributions, in addition to implications for option pricing. Andersen *et al.* (2002) provide in-sample specification tests as well as option pricing implications. Kaeck (2013) and Ignatieva *et al.* (2015) rely on the deviance information criterion, an in-sample Bayesian fit statistic developed in Spiegelhalter *et al.* (2002). Christoffersen *et al.* (2010) provide (in-sample) QQ plots as well scatter plots of variance level changes, and conclude that affine variance processes are rejected by the data. Li *et al.* (2008) apply (in-sample) kernel density plots, QQ plots and Kolmogorov-Smirnov (KS) tests to in-sample model residuals. Kou *et al.* (2013) also use KS tests and QQ plots in addition to comparing model autocorrelation functions to those observed in the data. Szerszen (2009) provides QQ plots and Value-at-Risk (VaR) specification tests based on in-sample parameters.

We test a variety of stock return models that have been proposed in the literature. Starting with the affine stochastic variance model proposed by Heston (1993), many extensions have been developed to model different features of the data. One area of research has focused on Poisson jump models, such as Bates (1996), or extensions to double-jumps

as in Duffie *et al.* (2000) and Eraker *et al.* (2003). Intuitively, such models allow for occasional spikes in the data (for instance the market crash of October 1987), which are captured by a finite-activity jump process. Variations of these models alter the jump size distribution or introduce time-varying jump intensities (see Kou, 2002, Pan, 2002 or Kaeck, 2013). More recently, a number of studies have introduced models based on infinite-activity Lévy processes. Li *et al.* (2008), Szerszen (2009), Bates (2012) and Ornthanalai (2014) provide evidence that suggests that such a modeling approach can be advantageous.[2] Another strand of the literature studies multifactor variance specifications, as these support more erratic variance movements (see Chernov *et al.*, 2003 or Kaeck and Alexander, 2012). The literature also presents convincing evidence in favor of non-affine variance dynamics (see Jones, 2003, Christoffersen *et al.*, 2010, Mijatovic and Schneider, 2014, or Ignatieva *et al.*, 2015), albeit often at the cost of tractability as these models do not allow for closed-form characteristic functions. Finally, discrete-time GARCH models as well as discrete-time stochastic volatility specifications provide alternative modeling frameworks; for recent surveys, we refer to Bauwens *et al.* (2006) and Andersen *et al.* (2009).

Our main data set comprises of daily observations of index returns. Another interesting strand of the literature, see e.g., Stroud and Johannes (2014) and Bates (2016), discusses modeling intradaily high-frequency return data. Analyzing high-frequency returns for a large set of models is beyond the scope of our paper. Data generating processes for higher frequency data, such as 5-minute returns, are substantially more complex because of well known intradaily trading patterns such as volatility seasonality and other market microstructure effects. These differences complicate model comparisons between different data frequencies. In addition, transition densities over more than one time step do not exist in closed form for most of the models used in our paper. Using high-frequency returns would also increase the computational burden which is already substantial given

[2] Lee and Hannig (2010) and Aït-Sahalia and Jacod (2011) propose statistical tests to distinguish between finite and infinite-activity jumps in high-frequency data.

the large number of models we consider.[3] While our tests focus on daily observations, to integrate the benefits of high-frequency data we follow Maneesoonthorn *et al.* (2017) and use intradaily realized volatility measures for model estimation. This approach reduces the computational burden compared to using intradaily returns directly and mitigates issues with comparing model forecasts at different observation frequencies.

3 Model Specifications

For the first model category, we assume that the log asset price $s_t = \ln S_t$ follows a jump-diffusion process with stochastic variance and a stochastic mean reversion level as proposed by Duffie *et al.* (2000), Egloff *et al.* (2010) and others:

$$ds_t = \left(\mu_c - \tfrac{1}{2}v_t - \lambda_t \bar{k}\right) dt + \rho_v \sqrt{v_t} dW_t^v + \sqrt{1-\rho_v^2}\sqrt{v_t}\, dW_t^s + \xi_t dN_t \quad (1)$$

$$dv_t = \kappa_v \left(m_t - v_t\right) dt + \sigma_v v_t^{\gamma} dW_t^v \quad (2)$$

$$dm_t = \kappa_m \left(\theta_m - m_t\right) dt + \sigma_m m_t^{\gamma} dW_t^m, \quad (3)$$

where μ_c is the drift and v_t denotes the stochastic variance process with speed of mean reversion κ_v and volatility parameter σ_v. The stochastic mean-reversion level m_t is governed by the speed of mean reversion κ_m, the long-term mean-reversion level θ_m and the diffusion parameter σ_m. All three Brownian motion processes W^v, W^s and W^m are uncorrelated, and as a consequence ρ_v determines the correlation between variance innovations and returns. The parameter γ identifies the dependence of the diffusion functions on the level of variance and long-term variance, respectively.[4] For $\gamma = \frac{1}{2}$ we obtain an extension of the standard affine specification of Heston (1993) (labeled A); for $\gamma = 1$ we

[3] Our calculations runs on a large computer cluster and each model estimation utilizes 15 parallel cores. Simple model specifications require about one day of computational time using daily return observations. Employing 5 minute frequency returns the algorithm would take more than 90 times longer.

[4] For simplicity, we use the same CEV parameter γ for both the variance and the long-term variance process.

have a continuous-time GARCH (henceforth CGARCH, see Nelson, 1990) process (G); and finally, if the parameter may take any value between one half and three halves, we obtain a general CEV variance model (C). Jump events occur at random times whenever increments in the Poisson counting process are equal to one, i.e. $dN_t = 1$. We assume that N has a state-dependent intensity $\lambda_t = \lambda_c + \lambda_v v_t$, where λ_c is the time-independent part of the jump intensity and λ_v measures the dependence of the jump probability on the current variance level. The iid jump size ξ_t is normally distributed with mean μ_s and standard deviation σ_s. Furthermore, we follow the standard convention that jump sizes are independent of all other stochastic variables. The jump compensator of this model is given by $\bar{k} = \exp\left[\mu_s + \frac{1}{2}\sigma_s^2\right] - 1$.

For alternative jump specifications, we follow Bates (2012) and focus on jump models driven by CMGY model dynamics (see Carr *et al.*, 2002) and assume that the log asset price dynamics in Equation (1) are replaced by

$$ds_t = \left(\mu_c - \tfrac{1}{2}v_t\right)dt + \rho_v\sqrt{v_t}dW_t^v + \sqrt{1-\rho_v^2}\sqrt{v_t}\,dL_t \tag{4}$$

where dL_t is the increment of a compensated Lévy process. The logarithm of the characteristic function $\Psi^{CMGY}(u,t) = \mathbb{E}\left[\exp\left\{uL_t^{CMGY}\right\}\right]$ of the generalized CGMY process of Carr *et al.* (2003) is given by

$$\ln\Psi^{CMGY}(u,t) = (\mu-\omega)ut + tV\left[w_n\frac{(G+u)^{Y_n} - G^{Y_n}}{Y_n(Y_n-1)G^{Y_n-2}} + (1-w_n)\frac{(M-u)^{Y_p} - M^{Y_p}}{Y_p(Y_p-1)M^{Y_p-2}}\right]$$

where ω is a normalizing constant, V is the variance per unit time and w_n determines the fraction of downward jumps. We further define

$$C_n = \frac{w_n V}{\Gamma(2-Y_n)\,G^{Y_n-2}} \quad \text{and} \quad C_p = \frac{(1-w_n)V}{\Gamma(2-Y_p)\,M^{Y_p-2}},$$

where $\Gamma(z)$ denotes the gamma function. With this definition, the parameter range is restricted to $C_n, C_p, G, M > 0$ and $Y_p, Y_n < 2$. For $Y_p, Y_n < 0$ the process has finite

activity, for $Y_p, Y_n < 1$ the process has finite variation. The model with $L_t = L_t^{CMGY}$ nests a wide range of models used in the finance literature; we follow Bates (2012) in using the parameter restrictions $Y_n = Y_p = 1$ for the double exponential (DEXP) jump model of Kou (2002), and $Y_n = Y_p = 0$ for the variance gamma (VG) model of Madan and Seneta (1990) for which a non time-changed version has been estimated in Li *et al.* (2008). The full CMGY is labeled YY whereas an extension to

$$\ln \Psi^{YYD}(u,t) = (1 - f_{jump})\tfrac{1}{2}(u^2 - u)t + f_{jump} \ln \Psi^{CMGY}(u,t)$$

for $f_{jump} \in [0,1]$ is called YYD, where D indicates an additional diffusive component. For further details such as Lévy densities or normalizing constants we refer to Bates (2012).

The asset price specifications introduced in this section allow us to distinguish between a wide range of models previously employed in the literature. The main model categories as far as the jump dynamics are concerned distinguish between either no jumps (such as in Heston, 1993), finite-activity jumps (Bates, 1996 or Duffie *et al.*, 2000) and infinite-activity jumps (Madan and Seneta, 1990, Carr *et al.*, 2002). The variance dynamics are subdivided into affine, GARCH and general non-affine CEV dynamics (Nelson, 1990, Jones, 2003 or Christoffersen *et al.*, 2010) for both one and two-factor variance models (Egloff *et al.*, 2010 or Bates, 2012). Our model setup differs substantially from previous research which has often focused on comparing models along a single-dimension. We also compare continuous-time specifications with popular discrete-time GARCH (henceforth DGARCH) models and introduce these in Section 7 of the paper for expositional ease. We provide an overview of the one-factor continuous-time models used in this paper in Table 4, two-factor versions of the models have an additional identifier MF. Further models are discussed in Section 8.

[Table 1 about here.]

4 Econometric Methodology

4.1 Model Estimation

Our econometric methodology builds on the maximum likelihood estimation proposed by Bates (2006). Under affine model specifications, let the Δ-step ahead conditional characteristic function of the asset return $r_{t+\Delta} = s_{t+\Delta} - s_t$ and latent state variables be given by

$$\begin{aligned}\Psi_t^{\mathcal{G}}(u_1, u_2, u_3) &\equiv \mathbb{E}\left[e^{u_1 r_{t+\Delta} + u_2 v_{t+\Delta} + u_3 m_{t+\Delta}} \middle| \mathcal{G}_t\right] \\ &= \exp\left\{\mathcal{A}(u_1, u_2, u_3, \Delta) + \mathcal{B}(u_1, u_2, u_3, \Delta)\, v_t + \mathcal{C}(u_1, u_2, u_3, \Delta)\, m_t\right\}\end{aligned}$$

where $\mathcal{G}_t = \sigma\left(\{S_\tau, v_\tau, m_\tau\} : \tau \leq t\right)$ is the σ-algebra (information) generated by both the asset price and the latent state variables and $u_1, u_2, u_3 \in \mathbb{C}$ (as long as well-defined). The functional form of the complex-valued functions $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ follows from Duffie *et al.* (2000). These functions satisfy ODEs that can be solved explicitly for one-factor affine variance specifications.[5] For two-factor models, we use the numerical algorithm developed in Bates (2012) to solve for $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. This approach yields approximations that are highly accurate for applications such as ours.

To describe the filtering algorithm, the joint characteristic function of the latent state variables (given the information generated by the asset returns) is defined as

$$\Lambda_t(u_2, u_3) \equiv \mathbb{E}\left[e^{u_2 v_t + u_3 m_t} \middle| \mathcal{Y}_t\right]$$

where $\mathcal{Y}_t = \sigma\left(\{S_\tau\} : \tau \leq t\right)$ is the information generated by observing the asset price only.

[5] We refer to Heston (1993), Bates (1996), Pan (2002) or Bates (2012) for the exact functional form of these functions.

By the law of iterated conditioning, it follows that

$$\begin{aligned}\Psi_t^{\mathcal{Y}}(u_1,u_2,u_3) &\equiv \mathbb{E}\left[e^{u_1 r_{t+\Delta}+u_2 v_{t+\Delta}+u_3 m_{t+\Delta}}\middle|\,\mathcal{Y}_t\right]\\ &= e^{\mathcal{A}(u_1,u_2,u_3,\Delta)}\,\Lambda_t(\mathcal{B}(u_1,u_2,u_3,\Delta),\mathcal{C}(u_1,u_2,u_3,\Delta)).\end{aligned}$$

As a result, standard Fourier inversion methods provide the probability of a return observation conditional on all past returns and the vector θ containing model parameters:

$$p\left(r_{t+\Delta}|\,\mathcal{Y}_t,\theta\right)=\frac{1}{2\pi}\int_{\mathbb{R}} e^{\imath u r_{t+\Delta}}\,\Psi_t^{\mathcal{Y}}(\imath u,0,0)\,du, \tag{5}$$

where $\imath$ is the imaginary unit.[6] We apply this numerical procedure to calculate the log-likelihood for the different model specifications. The last step in the filtering algorithm provides the update of Λ_t, and is given by[7]

$$\Lambda_{t+\Delta}(u_2,u_3)=\frac{1}{2\pi\,p\left(r_{t+\Delta}|\,\mathcal{Y}_t\right)}\int_{\mathbb{R}} e^{\imath u r_{t+\Delta}+\mathcal{A}(\imath u,u_2,u_3,\Delta)}\,\Lambda_t(\mathcal{B}(\imath u,u_2,u_3,\Delta),\mathcal{C}(\imath u,u_2,u_3,\Delta))\,du.$$

To start the procedure $\Lambda_0(u_2,u_3)$ is set to the unconditional characteristic function.[8]

Non-affine model specifications lack closed-form characteristic functions and hence the method described above cannot be directly applied. We estimate non-affine models by locally approximating them with an affine model specification. More specifically, we approximate the one-day ahead characteristic function by plugging $\sigma_v^a=\sigma_v\mathbb{E}\left[v_t^{\gamma-0.5}\middle|\,\mathcal{Y}_t\right]$ into the respective affine characteristic function. Compared to the standard Euler discretization applied in the literature (see Eraker *et al.*, 2003), such approximation is likely to be negligible over small time-steps because compared to an Euler discretization (which works well in practice, see Eraker *et al.*, 2003 or Li *et al.*, 2008), only part of the variance

[6] In the following we will drop the conditioning on the parameter vector from the notation for convenience. We will investigate the impact of parameter uncertainty on our results in Section 8.

[7] We use this characteristic function to implement a moment-matching procedure, see Bates (1996).

[8] As suggested in Bates (2006) the numerical stability of the integrals is improved by calculating the Fourier transform of a "shifted" density and then numerically invert this function. We refer to the appendix of Bates (2006) for more details on this procedure.

dynamics are kept constant. In Appendix A we provide simulation evidence to substantiate this claim and show that our estimation routine can accurately estimate affine and non-affine model parameters.

4.2 Model Confidence Sets

Our empirical results include a large number of models and hence pairwise model comparisons provide only limited insight. To allow for multiple comparisons, we employ the Model Confidence Set (MCS) procedure proposed by Hansen *et al.* (2011). A MCS is defined as a set that contains the best model(s) from a collection of competing models, say $\mathcal{M}^0$, with a user-specified level of confidence $(1-\alpha)$, where α denotes the significance level (typically 10% and 25%).[9] The *best* models are identified based on a user-specified criterion that quantifies the relative performance of the models. Various such criteria are introduced below. A desirable property of the MCS procedure is that it acknowledges the informativeness of the data. Whereas informative data lead to the MCS containing only a few models (or even just one model), less informative data result in the MCS containing more or potentially even all models. The MCS procedure does not make a statement about which model is the true model, as performance is assessed relative to other competing models.

To fix notation, let the competing models in $\mathcal{M}^0$ be indexed by $i = 1, \ldots, m_0$, with m_0 denoting the number of models in $\mathcal{M}^0$. A user-specified loss function $L_{i,t}$ measures the performance of each model i at time t, and the relative performance between model i and j is defined as $d_{ij,t} \equiv L_{i,t} - L_{j,t}$ for all $i, j \in \mathcal{M}^0$. The expected loss of model i is defined as $\mu_{ij} \equiv \mathrm{E}(d_{ij,t})$ according to which models are ranked, hence model i is preferred to j if $\mu_{ij} < 0$. The set of superior models is defined as $\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \quad \forall j \in \mathcal{M}^0\}$.

The objective of the MCS procedure is to determine $\mathcal{M}^*$. To estimate $\mathcal{M}^*$, candidate

[9] This interpretation is analogous to that of a classical confidence interval, hence the MCS contains the best models with a chosen confidence level. This procedure does not necessarily thereby identify one *best* model, as the MCS might consist of several models that are not statistically superior to one another.

models are evaluated using an equivalence test $\delta_{\mathcal{M}}$ and inferior models are subsequently removed from the initial model set based on the elimination rule $e_{\mathcal{M}}$. That is, a series of iterative hypothesis tests is performed, testing at each step the hypothesis

$$H_{0,\mathcal{M}} : \mu_{ij} = 0 \quad \forall i, j \in \mathcal{M}, \tag{6}$$

where $\mathcal{M} \subset \mathcal{M}^0$ and the alternative hypothesis, $H_{A,\mathcal{M}}$, is given by $\mu_{ij} \neq 0$ for some $i, j \in \mathcal{M}$. The equivalence test $\delta_{\mathcal{M}}$ is used to test $H_{0,\mathcal{M}}$ for all $\mathcal{M} \subset \mathcal{M}^0$. As long as the hypothesis is rejected, the elimination rule $e_{\mathcal{M}}$ is applied to determine the most inferior model of $\mathcal{M}$ which is then eliminated from $\mathcal{M}$ and by this means a sequence of sets $\mathcal{M}^0 = \mathcal{M}_1 \supset \mathcal{M}_2 \supset \cdots \supset \mathcal{M}_{m_0}$ is defined, where $\mathcal{M}_i = \{e_{\mathcal{M}_i}, \ldots, e_{\mathcal{M}_{m_0}}\}$. The procedure is repeated until $H_{0,\mathcal{M}}$ cannot be rejected any more. We call the set of all surviving models $\widehat{\mathcal{M}}^*_{1-\alpha}$, the model confidence set with confidence level $(1-\alpha)$.

Analogous to classical statistical inference, MCS p-values are defined as follows: $P_{H_{0,\mathcal{M}_i}}$ denotes the p-value related to hypothesis $H_{0,\mathcal{M}_i}$. The p-value $P_{H_{0,\mathcal{M}_i}}$ is calculated as $1 - F_i(t_i)$ for $F_i(t_i)$ being the cdf of the i-th test statistic t_i. A large value for the test statistic leads to small values for $P_{H_{0,\mathcal{M}_i}}$ with the interpretation that the hypothesis $H_{0,\mathcal{M}_i}$, that all models in $\mathcal{M}_i$ are equal, is likely to be statistically rejected. The MCS p-value for the model determined by elimination rule $e_{\mathcal{M}_j}$ is calculated using $\widehat{p}_{e_{\mathcal{M}_j}} = \max_{i \leq j} P_{H_{0,\mathcal{M}_i}}$. This makes it easy to determine whether a model belongs to $\widehat{\mathcal{M}}^*$ or not, as model i is an element of $\widehat{\mathcal{M}}^*_{1-\alpha}$ for a given significance level α if $\widehat{p}_{e_{\mathcal{M}_j}} \geq \alpha$. Therefore, the MCS-$p$-value is interpreted such that a model with a small p-value being unlikely to be a member of $\mathcal{M}^*$.

Specifying equivalence tests and elimination rules requires the choice of a loss function by which model performance is assessed. We use several different loss functions which are defined in Section 4.3 below. To test the performance of model i against alternative model specifications, Hansen *et al.* (2011) propose using a multiple t-statistics approach based on the test statistic $T_{R,\mathcal{M}} = \max_{i,j \in \mathcal{M}} |t_{ij}|$, with $t_{ij} = \bar{d}_{ij} / \sqrt{\widehat{var}(\bar{d}_{ij})}$ and with

$\bar{d}_{ij} = T^{-1} \sum_{t=1}^{T} d_{ij,t}$. Since the distribution of the test statistic is non-standard, a bootstrap algorithm is used to estimate the MCS p-values (see appendix of Hansen *et al.*, 2011). The natural elimination rule is then given as $e_{R,\mathcal{M}} = \max_{i \in \mathcal{M}} \sup_{j \in \mathcal{M}} |t_{ij}|$, i.e., in case of rejection of the null hypothesis, the rule eliminates the model that contributes most to the test statistic.

4.3 Loss Functions

4.3.1 Predictive Likelihood

The first loss function employed in this paper uses the predictive log-likelihood to compute the loss $L_{i,t}$ (see Amisano and Giacomini, 2007, Bao *et al.*, 2007 or Wilhelmsson, 2013). Let $f_{i,t}$ denote the predictive density of model i from time $t - \Delta$ to t. The relative performance between two models over time is then given by $\bar{d}_{ij} = T^{-1} \sum_t - \ln (f_{i,t}/f_{j,t})$, where T denotes the number of observations. The minus sign in front of the logarithm converts the log-likelihoods into a loss function, hence model i is preferred over model j if $\bar{d}_{ij}$ is negative.

4.3.2 Continuous-Ranked Probability Score

We use further loss functions proposed in Gneiting and Ranjan (2011), and in particular we focus on the continuous-ranked probability score (CRPS) which is defined as

$$CRPS(f, y) = \int_{\mathbb{R}} \left(F(z) - \mathbb{1}\{y \leq z\}\right)^2 dz,$$

where f is the forecasting density, F its corresponding cumulative distribution function and y denotes the realized outcome (the S&P 500 index return in our case).[10] Intuitively, this loss function measures the difference between the forecasting distribution of a model

[10] We use the standard notation $\mathbb{1}\{A\}$ for the indicator function which takes the value 1 if A is true and zero otherwise.

and the *optimal* forecast that would have resulted from perfect foresight. As shown by Laio and Tamea (2006), an equivalent representation of CRPS can be obtained as an integral over quantiles and is given by

$$CRPS(f, y) = 2 \int_0^1 \left(\mathbb{1}\left\{ y \leq F^{-1}(\alpha) \right\} - \alpha \right) \left(F^{-1}(\alpha) - y \right) d\alpha,$$

where $F^{-1}(\alpha)$ is the α-quantile of the forecasting distribution.

The advantage of CRPS over other scoring rules (such as the predictive likelihood) is that this loss function can be extended such that particular areas of the distribution function are weighted more heavily while ensuring propriety of the scoring rule. Gneiting and Ranjan (2011) propose weighted versions of CRPS defined as

$$CRPS_w(f, y) = 2 \int_0^1 \left(\mathbb{1}\left\{ y \leq F^{-1}(\alpha) \right\} - \alpha \right) \left(F^{-1}(\alpha) - y \right) w(\alpha)\, d\alpha,$$

where $w(\alpha)$ is a non-negative weight function on the unit interval. We follow Gneiting and Ranjan (2011) and, in addition to the un-weighted CRPS, use the following weight functions : $w(\alpha) = \alpha(1-\alpha)$ (center), $w(\alpha) = (2\alpha - 1)^2$ (tails), $w(\alpha) = \alpha^2$ (right tail) and $w(\alpha) = (1-\alpha)^2$ (left tail).

To fix the notation, the average CRSP of model i is defined as

$$\overline{CRPS}_{w,i} = \frac{1}{T} \sum_t CRPS_w(f_{i,t}, r_t).$$

Forecasts from densities $f_{i,t}$ are preferred over forecasts from densities $f_{j,t}$ if $\overline{CRPS}_{w,i} < \overline{CRPS}_{w,j}$. With $d_{ij,t} = CRPS_w(f_{i,t}, r_t) - CRPS_w(f_{j,t}, r_t)$ and $\hat{\sigma}_{ij}^2 = \frac{1}{T} \sum_t d_{ij,t}^2$ it can be shown that under the null hypothesis of vanishing expected scores, the test statistic

$$t_T = \sqrt{T} \left(\overline{CRPS}_{w,i} - \overline{CRPS}_{w,j} \right) \hat{\sigma}_{ij}^{-1}$$

asymptotically follows a standard normal distribution (assuming suitable regularity con-

ditions for which we refer to Gneiting and Ranjan, 2011). To use the CRPS loss function for calculation of model confidence sets, the average relative performance between model i and j is defined as $\bar{d}_{ij} = \overline{CRPS}_{w,i} - \overline{CRPS}_{w,j}$.

4.3.3 Asymmetric Value-at-Risk Loss Function

The third loss function we employ is proposed by González-Rivera *et al.* (2004) and has been designed specifically for testing the predictive power of models in the context of VaR estimation. The proposed loss function is given by

$$L_{i,t}^{VaR} = (r_t - \text{VaR}_{i,t}^{\alpha}) \times \left(\alpha - \mathbb{1}\left\{r_t < \text{VaR}_{i,t}^{\alpha}\right\}\right)$$

where $\text{VaR}_{i,t}^{\alpha}$ is the Value at Risk at significance level α for model i estimated at time $t - \Delta$ for a return horizon of Δ. The functional form of the loss function implies that deviations from VaR are weighted more heavily if $r_t < \text{VaR}_{i,t}^{\alpha}$, which is in line with the goal of avoiding large losses. The relative performance of two models i and j is given by $\bar{d}_{ij} = T^{-1} \sum_t \left(L_{i,t}^{VaR} - L_{j,t}^{VaR}\right)$.

5 Data

We employ daily log returns of the S&P 500 index for a period from January 2, 1987 until December 30, 2016. This data set overlaps with many previous studies such as Andersen *et al.* (2002) and Eraker *et al.* (2003). We separate the sample into an in-sample period from 1987 until 2006 and an out-of-sample period from 2007 until 2016. With tranquil and turbulent market regimes in both sub-samples, we have an ideal testing ground for the existence of jumps and the performance of alternative volatility specifications and their relative merits for out-of-sample forecasting. We report all parameters on a yearly basis and set $\Delta = \frac{1}{252}$. Table 5 provides summary statistics for the whole sample period,

as well as various sub-samples used in this paper.

[Table 2 about here.]

6 Empirical Results

In this section, we present in- and out-of-sample results for the one-factor jump-diffusion and Lévy-jump models. We focus on these models first for expositional ease, and discuss multi-factor variance models as well as discrete-time specifications in Section 7.

6.1 Parameter Estimates and In-Sample Performance

We report parameter estimates for the one-factor jump-diffusion models in Table 1. As most models have been extensively discussed in the literature, we provide only a short interpretation of our estimation results. For the standard Heston model (SV-A) we estimate a long-term variance of 0.029 (which translates into a yearly volatility level of 17.03%) and a *vol-of-vol* σ_v of 0.435. Eraker *et al.* (2003) for instance find 14.37% for the long-term volatility and 0.3614 for the volatility diffusion parameter in their less turbulent sample which ends before the dot-com bubble bursts. Our correlation estimate of -0.681 and the speed of mean reversion (5.449) are also in line with previous findings. The CGARCH and CEV model parameter estimates portray two patterns: first, a higher γ value leads to a lower speed of mean reversion and secondly, a slightly increased estimate of θ_v. Interestingly, for all model classes, the CEV parameter γ is statistically indistinguishable from the CGARCH specification ($\gamma = 1$). Jumps across all different model specifications occur less than once a year in the time-homogeneous jump specifications, but can occur slightly more frequently when the jump probability depends on the prevailing volatility regime. The jump parameter λ_c in all SVSJJ models is estimated to be to zero, and hence our results indicate that time-varying jump probabilities are an important feature of S&P 500

Table 1: In-sample parameter estimation results (One-factor jump diffusions).

This table reports the parameter estimation results for the one-factor jump-diffusion models. Parameter estimates correspond to annual units and decimal notation for percentages. The estimation period is from January 2, 1987 to December 30, 2006. The estimation is performed using an extension of the maximum likelihood method proposed in Bates (2006). For each parameter, we report the maximum likelihood estimates and the standard errors in parenthesis. Log-likelihood values for each model are given in the last row. For exact model definitions see Section (3).

Parameters	SV models			SVJ models			SVSJ models			SVSJJ models		
γ	0.500	1.000	1.000	0.500	1.000	1.115	0.500	1.000	0.995	0.500	1.000	1.095
			(0.047)			(0.082)			(0.087)			(0.093)
μ_c	0.041	0.055	0.057	0.048	0.063	0.065	0.049	0.062	0.063	0.051	0.064	0.068
	(0.006)	(0.016)	(0.011)	(0.031)	(0.023)	(0.024)	(0.022)	(0.023)	(0.023)	(0.029)	(0.026)	(0.023)
κ_v	5.449	2.549	2.610	3.343	1.949	1.554	3.632	2.020	2.102	3.586	2.143	1.722
	(0.722)	(0.685)	(0.810)	(0.638)	(0.642)	(0.725)	(0.664)	(0.655)	(0.761)	(0.667)	(0.674)	(0.771)
θ_v	0.029	0.036	0.035	0.026	0.031	0.034	0.025	0.030	0.029	0.025	0.029	0.031
	(0.002)	(0.007)	(0.006)	(0.003)	(0.008)	(0.011)	(0.003)	(0.007)	(0.007)	(0.003)	(0.006)	(0.009)
σ_v	0.435	2.963	2.948	0.309	2.355	3.624	0.309	2.274	2.242	0.307	2.292	3.266
	(0.016)	(0.101)	(0.420)	(0.021)	(0.147)	(1.075)	(0.022)	(0.151)	(0.734)	(0.022)	(0.153)	(1.126)
ρ_v	-0.681	-0.782	-0.774	-0.697	-0.773	-0.783	-0.700	-0.774	-0.772	-0.701	-0.774	-0.784
	(0.034)	(0.030)	(0.030)	(0.040)	(0.036)	(0.036)	(0.039)	(0.036)	(0.036)	(0.040)	(0.036)	(0.036)
λ_c				0.901	0.750	0.742				0.001	0.003	0.004
				(0.290)	(0.243)	(0.250)				(0.399)	(0.358)	(0.377)
λ_v							54.909	49.414	49.155	55.134	48.720	48.572
							(16.162)	(14.888)	(15.225)	(26.061)	(24.218)	(27.657)
μ_s				-0.028	-0.027	-0.029	-0.027	-0.026	-0.026	-0.027	-0.026	-0.025
				(0.018)	(0.009)	(0.010)	(0.008)	(0.008)	(0.009)	(0.014)	(0.013)	(0.014)
σ_s				0.059	0.056	0.053	0.054	0.050	0.050	0.054	0.050	0.048
				(0.008)	(0.007)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
LL	16722	16755	16755	16762	16783	16784	16769	16788	16788	16769	16788	16788

Table 2: In-sample parameter estimation results (One-factor Levy jump models).

This table reports the parameter estimation results for the one-factor Lévy jump models. Parameter estimates correspond to annual units and decimal notation for percentages. The estimation period is from January 2, 1987 to December 30, 2006. The estimation is performed using an extension of the maximum likelihood method proposed in Bates (2006). For each parameter, we report the maximum likelihood estimates and the standard errors in parenthesis. Log-likelihood values for each model are given in the last row. For exact model definitions see Section (3).

Parameters	SVYY models			SVDEXP models			SVVG models			SVYYD models		
γ	0.500	1.000	1.043	0.500	1.000	1.018	0.500	1.000	1.197	0.500	1.000	1.001
			(0.068)			(0.077)			(0.078)			(0.093)
μ_c	0.061	0.071	0.077	0.053	0.064	0.067	0.050	0.062	0.074	0.059	0.070	0.072
	(0.029)	(0.025)	(0.027)	(0.028)	(0.021)	(0.021)	(0.020)	(0.019)	(0.024)	(0.027)	(0.026)	(0.027)
κ_v	3.775	2.110	1.840	3.531	2.049	2.021	3.507	2.040	1.395	3.705	2.129	2.162
	(0.674)	(0.655)	(0.677)	(0.651)	(0.563)	(0.763)	(0.628)	(0.572)	(0.702)	(0.658)	(0.673)	(0.801)
θ_v	0.029	0.032	0.031	0.030	0.034	0.034	0.031	0.035	0.040	0.029	0.032	0.032
	(0.004)	(0.007)	(0.007)	(0.004)	(0.007)	(0.008)	(0.004)	(0.007)	(0.014)	(0.003)	(0.007)	(0.007)
σ_v	0.323	2.159	2.473	0.331	2.261	2.399	0.335	2.238	4.482	0.320	2.178	2.173
	(0.032)	(0.151)	(0.616)	(0.025)	(0.146)	(0.639)	(0.027)	(0.146)	(1.195)	(0.024)	(0.154)	(0.726)
ρ_v	-0.632	-0.711	-0.729	-0.640	-0.717	-0.717	-0.631	-0.711	-0.723	-0.632	-0.710	-0.713
	(0.054)	(0.048)	(0.042)	(0.041)	(0.037)	(0.039)	(0.044)	(0.041)	(0.046)	(0.042)	(0.048)	(0.046)
f_{jump}				0.289	0.301	0.297	0.338	0.339	0.365	0.998	1.000	0.997
				(0.077)	(0.081)	(0.081)	(0.104)	(0.102)	(0.127)	(0.000)	(0.000)	(1.977)
w_n	0.330	0.348	0.351	0.832	0.848	0.848	0.832	0.839	0.855	0.319	0.348	0.345
	(0.126)	(0.114)	(0.118)	(0.106)	(0.106)	(0.107)	(0.110)	(0.109)	(0.110)	(0.074)	(0.126)	(0.637)
G	3.761	3.944	6.479	27.500	28.000	28.000	14.188	15.306	13.363	4.996	3.838	3.990
	(3.018)	(2.833)	(6.914)	(6.045)	(6.190)	(5.875)	(5.759)	(5.816)	(5.809)	(0.002)	(3.399)	(3.952)
Y_n	1.297	1.368	1.637							1.122	1.379	1.399
	(0.450)	(0.460)	(0.276)							(0.055)	(0.473)	(0.514)
Y_p	1.858	1.837	1.809							1.855	1.812	1.804
	(0.058)	(0.075)	(0.088)							(0.055)	(0.086)	(0.615)
LL	16779	16797	16797	16771	16790	16790	16772	16791	16791	16780	16797	16797

index returns. These results confirm earlier evidence in Bates (2006) and Christoffersen *et al.* (2012). Jump sizes are relatively stable across different specifications with average means of -3% and a standard deviation between 5% and 6%.

The likelihood values at the optimal parameter set indicate that jump models substantially improve the in-sample performance, for instance we find that the log-likelihood increases by a value of between 30 to 40 from SV to SVJ. Time-varying jump probabilities provide further improvements of the log-likelihood, especially in the affine model specification. Consistent with the low parameter estimates for λ_c we find little evidence of an improvement of SVSJJ over SVSJ. The second very consistent result is that the CGARCH models clearly outperform affine models, whereas a free CEV parameter has only a minor effect on the performance measure.

We report the parameter estimates for the one-factor Lévy-jump models in Table 2. For the parameters that govern the stochastic variance process we find similar patterns to those for the one-factor jump-diffusion models. Long-term volatility estimates are quite stable and vary between 17 to 20%. Correlation estimates vary between -0.632 and -0.729 and are increasing slightly with an increasing γ. Estimates for κ_v and θ_v have a similar order of magnitude, and decrease and increase respectively with an increasing γ parameter. All affine versions of the one-factor Lévy-jump models (SVYY, SVDEXP, SVVG, SVYYD with $\gamma = 0.5$) are also estimated in Bates (2012) and the reported variance parameters are in line with our estimates: long-term volatility ($\sqrt{\theta_v}$) varies between 15.3 and 17.4%, mean-reversion speed (κ) varies between 3.961 and 8.318, and correlation (ρ) varies between -0.541 and -0.674. Also the parameter estimates for the Lévy-jump models given in Bates (2012) are of the same order of magnitude as ours and show the same structural behavior across model specification. Differences in the estimates can be explained by the different sample periods used in the papers. The weighting parameter w_n varies between 0.49 and 0.88, and our estimates vary between 0.32 and 0.83. The parameter f_{jump} which gives the proportion of variance that is driven by the Levy-jump part varies from 0.253 to 0.436, whereas we find values between 0.289 and 0.338. For the

SVYYD model we find that f_{jump} is close to the boundary value 1, which implies that the SVYYD model reduces to SVYY. Therefore, our data does not support an additional diffusion component for this model specification. This finding is confirmed by the log-likelihood values for the SVYYD and the SVYY model, which remain very close even for the different γ specifications.[11]

In terms of in-sample log-likelihood values, we obtain similar model rankings to those in Bates (2012). In particular, we find that the performance of DEXP models is similar to SVVG models and these are outperformed by SVYY model. The additional distribution component in the SVYYD model does not lead to further fit improvements, as log-likelihood values remain very close to the SVYY model.

We briefly discuss the in-sample performance of the models. Let $\mathcal{L}^m_{tT}$ be the log-likelihood of model m between time t and T. Then $\mathcal{R}^m_h = (h-t)^{-1}\left(\mathcal{L}^m_{th} - \mathcal{L}^b_{th}\right)$ for $h = t, \ldots, T$ defines the sequential normalized difference between the log-likelihood function of model m and the benchmark model b between t and $h > t$. In Figure 1 we compare these relative likelihood sequences over the in-sample period as suggested by Johannes *et al.* (2009), with SV-A as our benchmark. From the sequential likelihood ratios it is evident that the severe market shock of 1987 plays a crucial role in distinguishing different specifications, indeed all models cope with the large -23% return observed on October 19, 1987 far better than the affine SV-A model (see also the discussion in Eraker *et al.*, 2003). Jump models are slightly more successful during this extended period of market turmoil; simple non-linear variance models however also fare relatively well.

[Figure 1 about here.]

Overall jumps improve the likelihood ratios substantially and for the in-sample period, accounting for these is more important then the choice of variance dynamics. This is evident from the fact that affine jump models out-perform non-affine pure stochastic

[11] We have first estimated all Lévy models as in Bates (2012). Since our shorter sample period has significantly fewer large positive return outliers we found M to be unstable in the estimation and fixed the value to estimates in Bates (2012). All empirical results are robust as to whether M is fixed or estimated.

volatility models. Levy models provide further improvements over jump-diffusion specifications, and roughly half of the difference between jump vs non-jump models results from the October 1987 period.

6.2 Out-of-Sample Forecasting Performance – Log-Likelihood

In Figure 2, we present sequential likelihood ratios for the out-of-sample period which are calculated using Equation (5), fixing structural parameters to those estimated during the estimation window (as in Eraker, 2004).[12] Interestingly, this figure highlights a striking difference from our in-sample results, as the simple SV-G model outperforms all other specifications by roughly two log-likelihood points per year.

[Figure 2 about here.]

Jump models, although performing well in-sample, do not exhibit major improvements even over the simple affine stochastic volatility model SV-A. In addition, we find that the underperformance of jump models is gradually accumulated over the out-of-sample period rather than being the result of a single outlier. By contrast, the excess likelihood of non-linear variance models is accumulated predominantly during the outbreak of the financial crisis in 2008. We return to this finding further below.

[Table 3 about here.]

To add statistical rigor to our graphical results, in Table 6 we report model confidence set estimations using the out-of-sample negative log-likelihood as a loss function. We focus on affine and CGARCH models, and remove models of the CEV and SVSJJ class as their parameters (and out-of-sample results) are indistinguishable from other model specifications.[13] For the model confidence set estimation, we choose the block length of the bootstrap as follows. For each model, we estimate simple autoregressive (AR) models and determine the optimal lag length according to AIC and BIC fit criteria. We then

[12] We provide results for other out-of-sample designs in Section 8.
[13] Results for these models are available from the authors upon request.

select the bootstrap block length equal to the maximum lag length of all models in $\mathcal{M}_0$. It is evident from these results that the difference in out-of-sample likelihood between SV-G and all other specifications is statistically significant. We find that the MCS consists of the SV-A and SV-G models at the 25% level, which provides strong evidence in favor of simple stochastic volatility models. The first models eliminated from the initial model set are affine jump specifications. After this, jump models with CGARCH variance dynamics are excluded, and interestingly we find virtually no difference between the performance of finite and infinite-activity jump models. Our results confirm statistically the superiority of the simple CGARCH volatility specification and the fact that the MCS includes only two models can be interpreted as strong evidence that the out-of-sample period is informative with regard to the different model features. These findings also provide the first evidence that the choice of volatility dynamics is more important than modeling jumps.

[Table 4 about here.]

The graphical analysis in Figure 2 indicates that there may be two distinct regimes during the out-of-sample period: a first turbulent regime during the international financial crisis (2007-2009), and a second more stable regime until the end of the sample period (2010-2016). As it appears that most of the outperformance of the SV-A and SV-G models stems from the credit crisis period, we rerun the model confidence set estimation for both sub-periods separately to understand how the model ranking is affected by different market environments. Results in Panel A of Table 7 confirm that the CGARCH model class performs significantly better during the crisis period, and the model confidence set at the 25% level consists of all CGARCH specifications (with or without jumps), whereas all affine models perform weakly. This confirms the graphical findings that the variance dynamics are very important for adequately modeling market crashes. In Panel B, we focus on the calmer sub-period and find that the model confidence set at the 10% level consists only of the two stochastic volatility models, with insignificant performance differences between SV-A and SV-G. Taken together, the out-of-sample log-likelihood tests provide evidence that simple stochastic volatility models outperform more advanced

jump specifications and that the dynamics of the variance process matter, particularly during turbulent market regimes.

6.3 Out-of-Sample Forecasting Performance – Continuous Ranked Probability Score

We now provide out-of-sample results for a loss function that focuses on the forecasting performance of alternative models. In addition, we aim to test whether jump specifications provide superior performance in forecasting tail events, as one advantage of jump models is that they provide additional flexibility to fit the tails of the return distribution. To this end, we follow the framework of Gneiting and Ranjan (2011) and base our assessment of the forecasting performance on the continuous ranked probability score (for formal definitions, see Section 4.3).

[Table 5 about here.]

We report empirical results for the unweighted CRPS tests in Table 8.[14] The best performing model is SV-G, in line with the empirical results for the log-likelihood loss function above. In pairwise comparisons, this specification outperforms all other competing models at all conventional significance levels. The lowest absolute pairwise t-statistic results from the comparison with SV-A (with a t-statistic of -2.44). Given the large number of alternative models, this finding presents very strong support for non-linear variance dynamics. Furthermore, we find that diffusion models significantly outperform their jump extensions out-of-sample. In the affine model class the simple SV-A model outperforms all affine jump extensions, with the lowest significance level arising for a t-statistic of -3.28 for the comparison with the variance-gamma jump model. The same finding can be seen for the CGARCH model class where the smallest significance level results from the comparison between SV-G and SVJ-G (with a t-statistic of -3.51). Overall, CGARCH models offer a significant and very consistent improvement over affine models, with t-statistics

[14]For expositional ease we do not report the result for SVDEXP and SVSJ in these tables as they do not provide additional insights. We nevertheless include them in the model confidence set estimations below to ensure all empirical results are based on the same initial model set $\mathcal{M}_0$.

ranging from 2.44 to 3.51 when comparing the same jump specification with either an affine or CGARCH-type variance process.

[Table 6 about here.]

We restrict the detailed discussion of CRPS results for alternative weight functions to the left tail of the return distribution as this part of the distribution is most interesting for financial applications such as VaR. In addition, the left tail of the distribution of S&P 500 index returns benefits the most from the addition of jumps and hence weighting the left tail more heavily may uncover potential shortcomings of simple SV specifications. Our test results in Table 9 show that the model ranking is surprisingly little changed after altering the weight function. In particular, SV-G is still the overall best performing model and dominates all other specifications in pairwise model comparisons. However, jump models close the gap to simple SV models, and pairwise CRPS t-tests now indicate no statistically significant differences between the forecasting ability of competing model specifications. The empirical results for the forecasting tests with center, tails and right tail weight functions are available upon request, while for ease of exposition we restrict the discussion of these additional weight functions to the model confidence set estimation below.[15]

[Table 7 about here.]

We extend our previous findings and add MCS estimations to the pairwise model comparisons (see Table 10). Results for the different weight functions are in general supportive of the model rankings presented above and provide further strong evidence in favor of SV-G. Unsurprisingly, given the strong pairwise outperformance in Table 8, for all test statistics except for the left tail discussed above, SV-G is the only model in the 10% confidence set and it is also the best performing specification for all five test statistics. SV-G is particularly successful in the center and the right tail of the return distribution.

[15]Results for these tests, similar to Table 9, are available upon request.

It is also notable that the SV-A model provides a poorer performance compared to the log-likelihood loss function, where it was included in the MCS. Confirming our earlier findings, the model confidence set for the left tail includes all models, hence we are not able to distinguish between the forecasting performance in the left tail, at least as far as our out-of-sample period is concerned. Nevertheless, the SV-G model still performs best in this category, albeit at no conventional significance level.

[Table 8 about here.]

In Table 11, we report model confidence set results for the two distinct out-of-sample regimes (January 2007 to December 2009 and January 2010 to December 2016). The (unweighted) Gneiting-Ranjan tests confirm that during the financial crisis period models with CGARCH variance dynamics outperform affine specifications, whereas the addition of jumps does not lead to further improvements. The best-performing affine model is, as before, the simple diffusion specification, and this is the only affine model in the 25% model confidence set. By contrast, the second calmer period (January 2010 to December 2016) provides strong evidence that the SV-G and SV-A models dominate jump specifications (they are the only two models in the MCS at the 10% level). The results in Panel A and B of Table 11 provide supporting evidence that the forecasting performance in the left tail of the distribution is not improved with jumps of either finite or infinite activity as model confidence sets include all initial models $\mathcal{M}_0$. For completeness, we also report model confidence sets for alternative weight functions.

6.4 Out-of-Sample and Forecasting Performance – Berkowitz

Berkowitz (2001) proposes an alternative method for testing the forecasting performance, building on work from Diebold *et al.* (1998) and others. It is assumed that a forecasting model with density f is employed, whereas the true (unknown) density is given by p. It

can be shown that the density of the integral transform z, defined as

$$z = \int_{-\infty}^{y} f(z)\, dz = F(y),$$

is given by $p(F^{-1}(z))/f(F^{-1}(z))$. Therefore, under the null hypothesis that the forecasting model is equal to the true data-generating process, the variable z is uniformly distributed and $\tilde{z} = \Phi^{-1}(z)$ follows a standard normal distribution (Φ^{-1} denotes the inverse cumulative distribution function of a standard normal random variable). Furthermore, it can be shown that in a time-series framework, the realizations $\tilde{z}_t$ need to be iid. Berkowitz (2001) proposes to test this hypothesis using $\tilde{z}_t - \mu = \rho(\tilde{z}_{t-1} - \mu) + \varepsilon_t$ and the corresponding log-likelihood function $L(\mu, \sigma, \rho)$. This implies three possible tests, one for iid, one for independence and one for the joint hypothesis. The likelihood ratio test statistics are given by $LR_{ind} = -2\left[L(\hat{\mu}, \hat{\sigma}, 0) - L(\hat{\mu}, \hat{\sigma}, \hat{\rho})\right]$, $LR_{iid} = -2\left[L(0, 1, 0) - L(\hat{\mu}, \hat{\sigma}, 0)\right]$ and $LR = -2\left[L(0, 1, 0) - L(\hat{\mu}, \hat{\sigma}, \hat{\rho})\right]$. The main advantage of this test procedure is that it provides an absolute test of the forecasting performance.

[Table 9 about here.]

Table 12 presents results for the LR statistic as this provides the most general analysis.[16] Overall, we find that all models are rejected by the data. This is unsurprising as most models struggle with explaining the returns during the onset of the financial crisis in 2008, a year for which the null hypothesis is rejected by all models. Similar to our findings for the relative forecasting performance, we find that the statistics for CGARCH models are, however, more than half of the value for affine models and hence provide additional confirmation of the superiority of CGARCH models during the financial crisis period. Interestingly, during all years other than 2008, for all of the models, the null hypothesis cannot be rejected at the 5% level. Modeling differences are merely relevant during the crisis period, a finding that reinforces the conclusions above.

[16] Results for other statistics are available upon request. The overall results are further divided into out-of-sample tests for each individual year in the out-of-sample period.

7 Multi-factor and Discrete-Time Models

In this section, we extend our analysis in various directions and discuss the performance of two-factor jump diffusion models (introduced in Section 3) as well as simple DGARCH models. To focus on our main findings, we report model comparisons with a representative subset of one-factor models, namely SV-A, SV-G, SVSJ-A, SVSJ-G, SVYYD-A and SVYYD-G.

7.1 Multi-factor Variance Models

Parameter estimates for two-factor jump-diffusion models are based on the same in-sample period from January 2, 1987 until December 29, 2006 and are reported in Table 3. Our main results can be summarized as follows. First, the stochastic process m_t is slowly mean-reverting (estimates for κ_m range between 0.516 and 1.408) and it exhibits a relatively low diffusive volatility parameter σ_m. Secondly, the addition of the time-varying mean reversion level significantly alters the dynamics of stochastic variance. The process v_t is now much faster mean-reverting to m_t than it is in one-factor models and it is also significantly more volatile. In the SV-A model class, for instance, the mean reversion speed κ_v for one- and two-factor models is 5.449 and 26.928 respectively, whereas estimates for σ_v increase from 0.435 to 0.633. A high value for κ_v implies that v_t varies erratically around the long-term variance m_t. Thirdly, we find that the estimate for γ is slightly higher than in one-factor models with values of between 1.057 and 1.176. This is likely due to the fact that variance itself moves more violently around m_t and a higher CEV parameter facilitates such fast-moving behavior. Jump parameter estimates are comparable to the one-factor specifications discussed above. Given our previous findings regarding the minor importance of jumps for out-of-sample forecasting, we refrain from extending the analysis to Lévy-jump models and restrict our results to jump-diffusions to capture jump-like behavior.

[Figure 3 about here.]

The left part of Figure 3 shows in-sample sequential likelihood ratios for all two-factor jump-diffusion models (using SV-A as benchmark model). The overall evolution of these statistics is comparable to that for one-factor models; in particular, we find that jump models out-perform simple diffusion specifications and non-affine stochastic variance models also provide further improvements. The right-hand graph in Figure 3 documents out-of-sample sequential likelihood ratios. It is evident that the start of the global financial crisis in 2008 is an important time period for distinguishing the performance of alternative models, and non-affine specifications perform substantially better than affine models during this market regime. Interestingly, affine multi-factor models are particularly unsuccessful at explaining S&P 500 index returns during the crisis period. A possible explanation for this finding is that while the variance process in affine two-factor models is more erratic, its rapid mean-reverting behavior forces m_t to drive the overall variance level. Since v_t in one-factor models is more volatile than m_t it is possible that affine two-factor models are less successful at modeling more substantial variance changes. Unreported results confirm that in the affine models, the spot variance of one-factor models exceeds the variance levels of two-factor models during the peak of the financial market crisis in 2008. CGARCH and general CEV models appear to suffer less from this shortcoming.

Table 3: In-sample parameter estimation results (Multi-factor jump diffusions).

This table reports the parameter estimation results for the two-factor jump-diffusion models. The estimation period is from 2 January 1987 to 29 December 2006. The estimation is performed using an extension of the maximum likelihood method proposed in Bates (2006). For each parameter, we report the maximum likelihood estimates and the standard errors in parenthesis. Log-likelihood values for each model are given in the last row.

Parameters	SV models			SVJ models			SVSJ models			SVSJJ models		
γ	0.500	1.000	1.175	0.500	1.000	1.176	0.500	1.000	1.057	0.500	1.000	1.127
			(0.035)			(0.041)			(0.040)			(0.075)
μ_c	0.045	0.038	0.060	0.038	0.047	0.059	0.046	0.048	0.051	0.046	0.048	0.053
	(0.002)	(0.027)	(0.026)	(0.021)	(0.028)	(0.026)	(0.024)	(0.022)	(0.024)	(0.027)	(0.027)	(0.021)
κ_v	26.928	16.078	11.402	8.017	9.052	8.400	8.544	9.366	9.287	8.451	9.467	9.285
	(2.793)	(1.820)	(1.521)	(1.419)	(1.486)	(1.588)	(1.486)	(1.466)	(1.539)	(1.391)	(1.508)	(1.547)
σ_v	0.633	4.186	6.670	0.349	2.856	5.501	0.348	2.802	3.488	0.345	2.829	4.547
	(0.028)	(0.152)	(0.724)	(0.027)	(0.198)	(0.715)	(0.027)	(0.202)	(0.586)	(0.027)	(0.205)	(1.307)
κ_m	1.408	1.022	0.946	0.574	0.534	0.516	0.602	0.593	0.586	0.628	0.601	0.570
	(0.171)	(0.117)	(0.125)	(0.221)	(0.182)	(0.197)	(0.210)	(0.161)	(0.188)	(0.170)	(0.205)	(0.202)
θ_m	0.018	0.016	0.014	0.019	0.017	0.016	0.018	0.016	0.016	0.018	0.016	0.016
	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
σ_m	0.228	1.727	3.765	0.194	1.794	4.156	0.190	1.821	2.358	0.196	1.821	3.210
	(0.028)	(0.247)	(0.798)	(0.052)	(0.413)	(1.431)	(0.045)	(0.348)	(0.609)	(0.034)	(0.426)	(1.350)
ρ_v	-0.651	-0.832	-0.857	-0.744	-0.861	-0.906	-0.747	-0.865	-0.880	-0.747	-0.864	-0.891
	(0.034)	(0.023)	(0.022)	(0.036)	(0.027)	(0.026)	(0.036)	(0.027)	(0.027)	(0.037)	(0.027)	(0.027)
λ_c				0.976	0.852	1.385				0.012	0.011	0.048
				(0.285)	(0.299)	(0.461)				(0.440)	(0.010)	(0.303)
λ_v							56.381	51.751	52.239	57.074	50.258	49.514
							(17.275)	(16.977)	(16.886)	(25.771)	(17.426)	(20.483)
μ_s				-0.028	-0.024	-0.008	-0.027	-0.026	-0.025	-0.026	-0.027	-0.025
				(0.012)	(0.020)	(0.007)	(0.012)	(0.001)	(0.011)	(0.004)	(0.002)	(0.002)
σ_s				0.062	0.062	0.041	0.059	0.054	0.053	0.060	0.054	0.053
				(0.002)	(0.007)	(0.002)	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)
LL	16722	16781	16789	16771	16812	16816	16777	16816	16818	16777	16816	16820

[Table 10 about here.]

Table 13 presents out-of-sample model confidence set estimates for the negative predictive log likelihood loss function. These results complement the graphical results presented earlier and further compare the model performance of the two-factor specifications with the most successful specification of each model class of Section 6. The MCS estimates confirm that SV-A and SV-G are the best performing one-factor models, and there are only three additional two-factor models in the 25%-level confidence set, namely MF-SV-G, MF-SVSJJ-G and MF-SVJ-G. The best performing model is MF-SV-G, followed by SV-G which exhibits a MCS p-value of 0.8147. Although slightly less extreme than in the case of the one-factor specifications, two-factor models do not benefit from adding additional jumps to capture large outliers either, and it is more important to account for non-linear variance dynamics as affine multi-factor models perform particularly poorly. These findings suggest that using a multi-factor model with non-affine variance dynamics provides a similar performance to a simpler one-factor non-affine specification. We therefore conclude that two-factor models do not add significant gains for our out-of-sample data set. However, we do not find any evidence that more complex models lead to a deterioration in performance either.

[Table 11 about here.]

Table 14 provides further out-of-sample results, using the Gneiting and Ranjan (2011) test procedure with weighted CRPS test statistics as our loss function. The results for an unweighted objective function, similarly to the predictive likelihood results, suggest that non-affine model dynamics are important and that non-affine two-factor models provide similar out-of-sample performance to simple SV-G and SV-A models. As before, there is substantially less evidence in multi-factor models that jumps have a negative effect on the forecasting performance, and all non-affine MF models are included in the 25% model confidence set. As before, it proves very difficult to distinguish between the forecasting performances in the left tail of the return distribution, where all models except for MF-

SV-A and MF-SVSJJ-A are included in the 25% MCS. Non-affine models are superior at forecasting the right tail, where affine models perform poorly. While most of the attention in the literature is devoted to the left tail of the return distribution, the right tail may be of particular interest to investors with short positions.

7.2 Discrete-Time GARCH Models

In order to compare our results to simpler DGARCH models, we estimate further specifications that have been found to perform well in the discrete-time literature.[17] Our benchmark model is given by a multi-factor GJR-GARCH model (see Glosten *et al.*, 1993), written in a form that explicitly highlights the long-term variance level:

$$r_{t+1} = \mu + \varepsilon_{t+1} = \mu + \sqrt{h_{t+1}} z_{t+1} + I_{t+1}\xi_{t+1} - \lambda\mu_j \tag{7}$$

$$h_{t+1} = q_{t+1} - \left(\alpha_h + \tfrac{1}{2}\gamma_h\right)(q_t + \psi) + \beta_h (h_t - q_t) + (\alpha_h + \gamma_h \mathbb{1}_{\varepsilon_t<0})\,\varepsilon_t^2 \tag{8}$$

$$q_{t+1} = q_q - \left(\alpha_q + \tfrac{1}{2}\gamma_q\right)(q_q + \psi) + \beta_q (q_t - q_q) + (\alpha_q + \gamma_q \mathbb{1}_{\varepsilon_t<0})\,\varepsilon_t^2 \tag{9}$$

where h_t is the diffusive variance, q_t is the long-term variance level, α_h and β_h are model parameters determining the speed of mean reversion and how quickly the variance changes in response to a return shock, and γ_h determines the leverage effect. The long-term variance itself follows a GJR specification with parameters q_q, β_q, α_q and γ_q. Jumps in the asset price process are driven by iid Bernoulli variables I_t with probability $P(I_t = 1) = \lambda$ and ξ_t is normally distributed with mean μ_j and standard deviation σ_j. The variance of the jump component is given by $\psi = \mathrm{Var}(I_t\xi_t) = \lambda\left(\mu_j^2 + \sigma_j^2\right) - \lambda^2\mu_j^2$. The error term z_t is iid with zero mean and unit variance, driven by either a normal distribution or a standardized Student-t distribution with degree of freedom parameter η.

The choice of this general discrete-time model is driven by several considerations. First, the model in its unrestricted form includes all the features studied for continuous-time

[17] See, e.g., Bauwens *et al.* (2006) and Engle and Ng (1993).

models, namely a two-factor variance process, jumps and fat-tailed (non-Gaussian) error term distributions. And secondly, the model allows us to study nested, more parsimonious model specifications to test which features of discrete-time models are important in out-of-sample exercises. We label the single factor models GJR-N and GJR-t, depending on the distribution of the error term. For these two specifications, we apply the restriction $\lambda = 0$ and $q_t = \bar{q}$ where $\bar{q}$ is a constant. The corresponding two-factor DGARCH models are labeled MF-GJR-N and MF-GJR-t. For models with Gaussian error terms we also include models with normally distributed jumps, and we add an additional J-identifier for these jump specifications.

[Table 12 about here.]

Table 15 reports parameter estimates for the single- and two-factor DGARCH models. Following the discrete-time literature, parameters are estimated on daily percentage returns $r^p_{t+\Delta} = 100 \times (s_{t+\Delta} - s_t)$ during the in-sample period from January 2, 1987 until December 29, 2006. For ease of comparison, we scale the log-likelihood at the optimal parameter set to be comparable with previously reported continuous-time models. For the sake of brevity, we do not discuss parameter estimates in detail; they are consistent overall with earlier results and values reported in the literature. Models with t-distributed error terms notably perform best in-sample, with large log-likelihood improvements over Gaussian models. Interestingly, single- and multi-factor models with normally distributed error terms and jumps are also outperformed by simpler models with fat-tailed error terms.

[Table 13 about here.]

[Table 14 about here.]

In Tables 16 and 17 we compare the DGARCH model performance to various continuous-time benchmark models. For the predictive log-likelihood loss function in Table 16, model confidence sets at both the 10% and 25% level are not affected by the addition of DGARCH models, and the outperformance of SV-G is still significant. This highlights the superiority

of continuous-time specifications over sophisticated DGARCH models. The best models from the DGARCH model class are GJR-t and MF-GJR-t models, and hence our results imply that the most important out-of-sample feature to consider is a fat-tailed error term. The Gneiting-Ranjan tests are summarized in Table 17 and also add further support to earlier findings. SV-G outperforms all other specifications and the 10% model confidence set is a singleton for four of the five weight functions (no weight, center, right tail). For modeling the left tail of the return distribution, we find that the 10% model confidence set includes all models, whereas the 25% set includes all but the MF-GJR-N specification. Driven by this finding, the results for the tail weight function provide evidence in favor of SV-G, with simple one-factor GJR models also providing adequate performance. To shed further light on the model ranking within the DGARCH class, we run a separate set of model confidence set estimations (unreported). These results confirm that GJR-t is the most successful discrete-time specification, being the only model in the 25% confidence set for the unweighted, center, right-tail and left-tail loss function.

We consider additional out-of-sample exercises using a Value-at-Risk loss function as described in Section 4.3. To economize on space, we relegate these additional results to Appendix B.

8 Additional Results

In this section we provide further empirical results that address a range of additional research questions: Are our results robust to using frequent updating of the structural parameters? What is the effect of parameter uncertainty on our forecasting exercise? Do additional data sources, such as high frequency data or derivatives data, confirm our earlier findings? Can we obtain further performance improvements by using a time varying mean in the return equation?

For these additional results we deviate from the maximum likelihood estimation method-

ology of Bates (1996) for three main reasons. First, the filtering algorithm based on characteristic functions is computationally prohibitive if the observed data are high-dimensional. Numerical integrations in higher dimensions (as in Equation (5)) would slow down the filter significantly, even if the data dimension is moderate.[18] Second, derivatives data are not straightforward to incorporate into our framework as they are usually non-linear functions of the state variables. And third, extending window estimations require frequent re-estimation of all models which is computationally infeasible given the large number of models and the long out-of-sample period we use.

To incorporate all extensions in a single estimation framework, as well as to provide additional robustness results by comparing our earlier findings to those obtained from a different estimation methodology, we rely on the SMC^2 algorithm of Chopin *et al.* (2013) and Fulop and Li (2013). SMC^2 can be regarded an extension of the batch importance sampler of Chopin (2002) which exploits the unbiasedness of likelihood estimates by particle filters (see Del Moral, 2004). SMC^2 methods perform Bayesian inference in state space models by combining Monte Carlo algorithms in both the state and parameter dimension to sequentially reweight parameter particles. To avoid particle impoverishment, the algorithm includes rejuvenation of particles through a resampling step and a MCMC update step. As a by-product of the estimation we obtain for each model $\mathcal{M}$ the predictive probability

$$p\left(r_{t+\Delta}|\mathcal{Y}_t,\mathcal{M}\right)=\int p\left(r_{t+\Delta}|\mathcal{Y}_t,\theta,\mathcal{M}\right)p\left(\theta|\mathcal{Y}_{1:t},\mathcal{M}\right)d\theta \tag{10}$$

which takes into account parameter uncertainty (by integrating over the posterior density of the structural parameters) and parameter learning (by updating $p\left(\theta|\mathcal{Y}_{1:t},\mathcal{M}\right)$ for every t).[19] Note that the information set $\mathcal{Y}_t$ can now include the VIX index and realized variance

[18] Bates (1996) (p. 912) writes: *"Second, AML has a "curse of dimensionality" originating in its use of numerical integration. It is best suited for a single data source; two data sources necessitate bivariate integration, while using higher-order data is probably infeasible. However, extensions to multiple latent variables appear possible."*

[19] Note the difference to Equation (5) which conditions on the parameter vector θ.

estimates besides return observations. Note further, that the predictive densities are out-of-sample and can be used to construct (log) Bayes factors for model comparison similar to our likelihood based out-of-sample model comparison in Section 6.[20]

Our implementation of the SMC^2 algorithm requires the design and implementation of fast and accurate particle filters to estimate the model likelihood. Efficient auxiliary particle filters for jump-diffusion models as well as Levy-jump models have been discussed in Johannes *et al.* (2009) or Fulop and Li (2013) to which we refer for further details. Bayesian estimations also require the specification of parameter priors and we use diffuse priors, in line with previous research.[21] SMC^2 algorithms are inherently parallel and benefit substantially from the use of GPU programming (see Lee *et al.*, 2010). Since we use a large number of models, we parallelize particle filters on the CPU and estimate all models on a large computer cluster. Parallelized estimations use 15 CPUs per model, resulting in estimation times between between one and several days for a single model, depending on model complexity and data.[22]

8.1 Parameter updating and parameter uncertainty

[Figure 4 about here.]

Our first goal is to study deviations from the fixed in- and out-of-sample periods used in earlier sections of the paper as well as to examine the extent to which parameter uncertainty affects our conclusions. Our out-of-sample design relied on the predictive density $p\left(r_{t+\Delta}|\mathcal{Y}_t, \hat{\theta}, \mathcal{M}\right)$ where $\hat{\theta}$ is the ML estimate of the structural parameter vector for the in-sample period. In contrast, the SMC^2 methodology allows us to update the parameter vector daily and integrate out parameter uncertainty. This may be particularly important for parameters affecting infrequent jumps for which credible intervals tend to

[20] As unconditional priors for the parameter vector at the start of the algorithm we use independent normal distributions. When necessary we truncate the distribution to the range of acceptable parameter values, e.g., the parameter ρ is drawn from a normal distribution truncated to the values between -1 and 1.

[21] Further details on the particle filter and prior specifications are available upon request.

[22] We perform our calculations on a large computer cluster equipped with Intel Xeon 2.6 GHz processors.

be larger then for diffusion parameters. In addition, each jump observation constitutes a substantial amount of additional information on the jump distribution. In line with our sequential log-likelihood diagnostics for the ML estimates, we base our model comparison on sequential log Bayes factors which are defined as

$$lBF_{it} = \log p\left(\mathcal{Y}_{1:t}|\mathcal{M}_i\right) - \log p\left(\mathcal{Y}_{1:t}|\mathcal{M}_B\right) \tag{11}$$

where $\mathcal{M}_i$ is the model under consideration and $\mathcal{M}_B$ is a common benchmark model. As in Section 6, we set $\mathcal{M}_B$ = SV-A. To interpret our results we follow the rule of thump put forward in Kass and Raftery (1995) and consider that the evidence against a model is positive if the log odds ratio is between 2 and 6, strong if it is between 6 and 10, and very strong if it is greater than 10.

Figure 4 presents log Bayes factors for the out-of-sample period (but now updating parameters daily) and confirms a range of earlier findings. We restrict the models to a subset for ease of presentation. First, affine jump models are out-performed by the simple SV diffusion model. And second, non-affine variance dynamics are very successful at the onset of the global financial market crisis in 2008 when all non-affine variance models provide substantial improvements over their affine counterparts. As before, log Bayes factors of non-affine models gradually decline from the end of 2008 and reach levels near zero towards the end of the sample in 2017. These results confirm that our earlier findings are not driven by our out-of-sample design or the choice of estimation methodology. Note that the only notable difference to our earlier results is that SV-A performs as well as SV-G when allowing for frequent parameter updates. We also find that the updating of parameters improves the performance of the non-affine VG jump model, although its affine counterpart performs quite poorly.

The right-hand graph in Figure 4 presents our results for two-factor extensions of the benchmark models. As before, the start of the global financial crisis in 2008 is an important time period for distinguishing the performance of alternative models, and non-

affine specifications perform substantially better than affine models during this market regime. We also confirm that affine multi-factor models are not particularly successful at the start of the crisis period. Compared to the one-factor specifications, however, the performance of non-affine two-factor models after the crisis is at par with SV-A and the model ranking (also within the affine two-factor model class) is very stable after the outburst of the crisis.

8.2 Additional data sources

[Figure 5 about here.]

We now explore the use of additional data sources to distinguish alternative model specifications. Two common sources are high-frequency returns and derivatives data which we discuss in turn. High frequency returns have become a popular source for non-parametric volatility estimators (see e.g. Barndorff-Nielsen and Shephard (2002), Creal (2008), Takahashi *et al.* (2009), or Hansen *et al.* (2012) among others) but have also been used to improve the estimation of parametric models. As the focus of our paper is on the daily predictive density of stock index returns, we follow Maneesoonthorn *et al.* (2017) and aggregate intradaily 5 minute returns into a realized variance measure which is used as a signal for the unobservable latent variance. More specifically, we assume that (a transformation of) realized variance is a second observation equation, in addition to the index returns, and given by

$$f(RV_t) = f(a + bv_t) + \varepsilon_t$$

where ε_t is a normally distributed error term. This specification follows the estimation methodology in Maneesoonthorn *et al.* (2017) using $f(x) = \log x$.[23] The main implication

[23]The additional transformation using f is not required for model estimation and one could choose $f(x) = x$. In order to avoids excessive weights on data during high-volatility periods and outliers in the data, we use a log transformation. In addition, using the continuous-time dynamics of our proposed models, it is possible to link the parameters a and b to structural parameters of the model. Instead we

of the extended data is that additional realized variance measures allow to pin down variance dynamics more easily due to the additional signal. We rely on realized variance measures from Oxford Man institute which is available from January 3, 2000 on which we start our estimation procedure. Compared to our earlier results, models are therefore estimated using a shorter time period but with additional data.

Alternatively, one may rely on derivatives market data, following Eraker (2004) and others. Similar to the approach of using realized variance measures in the estimation, financial derivatives are very informative on the latent state of the variance process and hence provide a valuable data source for distinguishing alternative models. One complication arising from dervatives data in our set-up is that non-affine models (as well as the multifactor models we consider) lack closed-form solutions to standard European options. To circumvent these issues we use the VIX index as an input to our estimation. The VIX index has the advantage of aggregating the information in the cross section of options, thereby reducing the dimensionality of the estimation. In addition, for the models considered in this paper, the squared VIX index can be shown to be a linear function of the latent variance (see Kaeck and Alexander, 2012) which implies that we can use a similar functional relationship as before:

$$f(VIX_t^2) = f(a + bv_t) + \varepsilon_t.$$

Note that the parameters a and b are model dependent and are functions of the risk-neutral parameters of the estimated model.

One advantage of our set-up is that we do not need impose a particular measure transformation. To demonstrate our approach, we take the SVJ model as an example. As Broadie *et al.* (2007) discuss, the risk-neutral measure allows for different jump parameters under the risk neutral measure $(\lambda^{\mathbb{Q}}, \mu_y^{\mathbb{Q}}, \sigma_y^{\mathbb{Q}})$, and in addition both variance drift param-

follow Maneesoonthorn *et al.* (2017) and estimate a and b as free parameters, implicitly assuming that the realized variance is a noisy signal for the unobservable model implied variance. This is due to the fact that, since realized variance is calculated intradaily and ignores over-night returns, the theoretical relationship holds only approximately in the data.

eters may be different under the risk-neutral measure $(\kappa_v^{\mathbb{Q}}, \theta_v^{\mathbb{Q}})$ as discussed in Cheridito *et al.* (2007). This implies that option data are only informative about the jump distributions under the risk-neutral measure and unless arbitrary assumptions on the relation between risk-neutral and real-world jumps are imposed, option prices are not informative on real world jumps. They are of course very informative on risk-neutral jumps, see Andersen *et al.* (2017). Since a and b are functions of all five model parameters, our approach aggregates these parameters without any loss of information.[24] Note, however, that the dynamics of the VIX index depends crucially on the real-world parameters and the assumed model dynamics and this additional information helps to pin down variance dynamics.

Log Bayes factors using RV data are presented in the left graph of Figure 5. These results confirm the superiority of non-affine dynamics during the onset of the crisis. Similarly, we find large improvements in non-affine models when using the VIX index as a signal for the latent variance. Compared to the big difference between affine and non-affine models, the differences within these model classes is small. Note that the VIX index provides a much stronger signal, as realized variance measures are more erratic (i.e. higher volatility for ε_t) than the VIX index. This results in much larger Bayes factors in the case of the VIX index which underline the strong out-performance for non-affine specifications. With both additional data sources, jump model specifications are of second-order importance.

8.3 Time varying mean

[Figure 6 about here.]

In this section, we discuss the performance of further model extensions and focus on the role of the expected return. In our specifications, expected log returns are given by $\mu_t = \mu_c - \frac{1}{2}v_t$. This may be overly restrictive and hence we investigate the performance of several

[24] For pricing European option contracts one would need to disentangle the effect of all parameters on a and b.

extensions of this base case. As a first straightforward extensions, we follow Bates (2006) and allow for an unrestricted linear dependence of the return on the variance regime: $\mu_t = \mu_c + b_v v_t$ where b_v is an additional model parameter. We find that these extensions yield almost identical results to previously reported evidence, and b_v is not significantly different from $-1/2$ in all model specifications. Second, we estimate extensions for which the drift is given by $\mu_t = \mu_c + \sum_i \beta_i X_{it}$ where X denotes additional data. We use the realized variance and a jump realization measure based on the difference between realized and bi-power variation (both averaged over the last 22 trading days) but find no evidence that these models improve the performance of our simple constant drift specification. Finally, we allow for time-varying drift specifications given by

$$d\mu_t = \kappa_\mu \left(\theta_\mu - \mu_t\right) dt + \sigma_\mu dW_t^\mu \tag{12}$$

allowing for a time-variation independent of other latent state variables. We report the estimation results for these general stochastic drift models in what follows.

Figure 6 presents Bayes factors for the stochastic drift specifications for which we add SM (stochastic mean) to the model identifier. On the left hand side of Figure 6, the general patterns confirm the findings discussed above. The addition of a stochastic drift does not change our conclusions, but helps to improve the model performance of our proposed specifications further. This is most clearly seen in Figure 6 (b) which includes specifications with constant and stochastic drift for which we see significant improvements espcially for longer out-of-sample periods with changing market conditions.[25]

9 Conclusion

This paper studies the out-of-sample performance of several popular time-series models for S&P 500 index returns. We use an in-sample data set from 1987 to 2006 for model

[25] We would like to thank an anonymous referee for suggesting this additional analysis.

estimation and test how well alternative models fare in explaining index returns during an out-of-sample period starting in 2007. We test a plethora of models, including finite- and infinite-activity jumps, non-affine variance and multi-factor variance specifications, in discrete- and continuous-time. Model specification tests include likelihood-based statistics and weighted and unweighted continuous-ranked probability scores which are combined with the model confidence set procedure of Hansen *et al.* (2011).

We find that despite the highly turbulent out-of-sample market regime, simple stochastic volatility diffusions outperform more advanced jump specifications. The most important model feature is the non-affinity of the variance process; other model features are found to provide no further improvement during the out-of-sample period of this paper. Furthermore, we find that jump-diffusion models with a constant intensity parameter are misspecified for out-of-sample prediction. Our results in combination with findings in Santa-Clara and Yan (2010) suggest that improving the modeling of the time-variation in jump distributions and jump intensities are promising directions for future research.

References

Aït-Sahalia, Y. and Jacod, J. (2011). Testing whether jumps have finite or infinite activity. *Annals of Statistics*, **39**(3), 1689–1719.

Amisano, G. and Giacomini, R. (2007). Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business & Economic Statistics*, **25**(2), 177–190.

Andersen, T., Benzoni, L., and Lund, J. (2002). An Empirical Investigation of Continuous-Time Equity Return Models. *The Journal of Finance*, **57**(3), 1239–1284.

Andersen, T. G., Davis, R. A., Kreiß, J. P., and Mikosch, T. (2009). *Handbook of financial time series*. Springer Berlin Heidelberg.

Andersen, T. G., Fusari, N., and Todorov, V. (2017). Short-term market risks implied by weekly options. *The Journal of Finance*.

Bao, Y., Lee, T., and Saltoglu, B. (2007). Comparing Density Forecast Models. *Journal of Forecasting*, **26**, 203–225.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Econometric analysis of realized volatility and its use in estimating stochastic volatility models. *Journal of the Royal Statistical Society B*, **64**(2), 253–280.

Bates, D. S. (1996). Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutsche Mark Options. *Review of Financial Studies*, **9**(1), 69–107.

Bates, D. S. (2006). Maximum Likelihood Estimation of Latent Affine Processes. *Review of Financial Studies*, **19**(3), 909–965.

Bates, D. S. (2012). U.S. stock market crash risk, 1926-2010. *Journal of Financial Economics*, **105**(2), 229–259.

Bates, D. S. (2016). How crashes develop: intradaily volatility and crash evolution. *Working Paper*.

Bauwens, L., Laurent, S., and Rombouts, J. V. K. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, **21**(1), 79–109.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, **19**(4), 465–474.

Broadie, M., Chernov, M., and Johannes, M. (2007). Model specification and risk premia: Evidence from futures options. *The Journal of Finance*, **62**(3), 1453–1490.

Carr, P., Geman, H., Madan, D. B., and Yor, M. (2002). The Fine Structure of Asset Returns: An Empirical Investigation. *The Journal of Business*, **75**(2), 305–333.

Carr, P., Geman, H., Madan, D., and Yor, M. (2003). Stochastic volatility for Lévy processes. *Mathematical Finance*, **13**(3), 345–382.

Cheridito, P., Filipović, D., and Kimmel, R. L. (2007). Market price of risk specifications for affine models: Theory and evidence. *Journal of Financial Economics*, **83**(1), 123–170.

Chernov, M., Gallant, A. R., Ghysels, E., and Tauchen, G. (2003). Alternative models for stock price dynamics. *Journal of Econometrics*, **116**(1-2), 225–257.

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, **89**(3), 539–552.

Chopin, N., Jacob, P. E., and Papaspiliopoulos, O. (2013). SMC2: An efficient algorithm for sequential analysis of state space models. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **75**(3), 397–426.

Christoffersen, P., Jacobs, K., and Mimouni, K. (2010). Volatility Dynamics for the S&P500: Evidence from Realized Volatility, Daily Returns, and Option Prices. *Review of Financial Studies*, **23**(8), 3141–3189.

Christoffersen, P., Jacobs, K., and Ornthanalai, C. (2012). Dynamic jump intensities and risk premiums: Evidence from S&P500 returns and options. *Journal of Financial Economics*, **106**(3), 447–472.

Creal, D. D. (2008). Analysis of filtering and smoothing algorithms for Lévy-driven stochastic volatility models. *Computational Statistics & Data Analysis*, **52**, 2863–2876.

Del Moral, P. (2004). Feynman-kac formulae. In *Feynman-Kac Formulae*, pages 47–93. Springer.

Diebold, F. C., Gunther, T. A., and Tay, A. S. (1998). Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review*, **39**(4), 863–883.

Duffie, D., Pan, J., and Singleton, K. (2000). Transform Analysis and Asset Pricing for Affine Jump-Diffusions. *Econometrica*, **68**(6), 1343–1376.

Egloff, D., Leippold, M., and Wu, L. (2010). The Term Structure of Variance Swap Rates and Optimal Variance Swap Investments. *Journal of Financial and Quantitative Analysis*, **45**(05), 1279–1310.

Engle, R. F. and Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance*, **48**(5), 1749–1778.

Eraker, B. (2004). Do stock prices and volatility jump? Reconciling evidence from spot and option prices. *The Journal of Finance*, **59**(3), 1–37.

Eraker, B., Johannes, M., and Polson, N. (2003). The Impact of Jumps in Volatility and Returns. *The Journal of Finance*, **58**(3), 1269–1300.

Fulop, A. and Li, J. (2013). Efficient learning via simulation: A marginalized resample-move approach. *Journal of Econometrics*, **176**(2), 146–161.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *Journal of Finance*, **48**(5), 1779–1801.

Gneiting, T. and Ranjan, R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, **29**(3), 411–422.

González-Rivera, G., Lee, T.-H., and Mishra, S. (2004). Forecasting volatility: A reality check based on option pricing, utility function, value-at-risk, and predictive likelihood. *International Journal of Forecasting*, **20**(4), 629–645.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The Model Confidence Set. *Econometrica*, **79**(2), 453–497.

Hansen, P. R., Huang, Z., and Shek, H. H. (2012). Realized GARCH : A Joint Model for Returns and Realized Measures of Volatility. *Journal of Applied Econometrics*, **27**, 877–906.

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies*, **6**(2), 327–343.

Ignatieva, K., Rodrigues, P., and Seeger, N. (2015). Empirical Analysis of Affine Versus Nonaffine Variance Specifications in Jump-Diffusion Models for Equity Indices. *Journal of Business & Economic Statistics*, **33**(1), 68–75.

Johannes, M. S., Polson, N. G., and Stroud, J. R. (2009). Optimal Filtering of Jump Diffusions: Extracting Latent States from Asset Prices. *Review of Financial Studies*, **22**(7), 2759–2799.

Jones, C. S. (2003). The dynamics of stochastic volatility: evidence from underlying and options markets. *Journal of Econometrics*, **116**(1-2), 181–224.

Kaeck, A. (2013). Asymmetry in the jump-size distribution of the S&P 500: Evidence from equity and option markets. *Journal of Economic Dynamics & Control*, **37**(9), 1872–1888.

Kaeck, A. and Alexander, C. (2012). Volatility dynamics for the S&P 500: Further evidence from non-affine, multi-factor jump diffusions. *Journal of Banking and Finance*, **36**(11), 3110–3121.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kou, S., Yu, C., and Zhong, H. (2013). Jumps in Equity Index Returns Before and During the Recent Financial Crisis: A Bayesian Analysis. *Working Paper*.

Kou, S. G. (2002). A Jump-Diffusion Model for Option Pricing. *Management Science*, **48**(8), 1086–1101.

Laio, F. and Tamea, S. (2006). Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrology and Earth System Sciences Discussions*, **3**(4), 2145–2173.

Lee, A., Yau, C., Giles, M. B., Doucet, A., and Holmes, C. C. (2010). On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. *Journal of computational and graphical statistics*, **19**(4), 769–789.

Lee, S. S. and Hannig, J. (2010). Detecting jumps from Lévy jump diffusion processes. *Journal of Financial Economics*, **96**(2), 271–290.

Li, H., Wells, M. T., and Yu, C. L. (2008). A Bayesian Analysis of Return Dynamics with Lévy Jumps. *Review of Financial Studies*, **21**(5), 2345–2378.

Madan, D. B. and Seneta, E. (1990). The Variance Gamma (V.G.) Model for Share Market Returns. *The Journal of Business*, **63**(4), 511–524.

Maneesoonthorn, W., Forbes, S., and Martin, G. M. (2017). Inference on Self-Exciting Jumps in Prices and Volatility Using High-Frequency Measures. *Journal of Applied Econometrics*, **32**, 504–532.

Mijatovic, A. and Schneider, P. (2014). Empirical Asset Pricing with Nonlinear Risk Premia. *Journal of Financial Econometrics*, **12**(3), 479–506.

Nelson, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics*, **45**(1-2), 7–38.

Ornthanalai, C. (2014). Levy jump risk: Evidence from options and returns. *Journal of Financial Economics*, **112**, 69–90.

Pan, J. (2002). The jump-risk premia implicit in options: evidence from an integrated time-series study. *Journal of Financial Economics*, **63**(1), 3–50.

Santa-Clara, P. and Yan, S. (2010). Crashes, Volatility, and the Equity Premium: Lessons from S&P 500 Options. *Review of Economics and Statistics*, **92**(2), 435–451.

Shackleton, M. B., Taylor, S. J., and Yu, P. (2010). A multi-horizon comparison of density forecasts for the S&P 500 using index returns and option prices. *Journal of Banking and Finance*, **34**(11), 2678–2693.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.

Stroud, J. R. and Johannes, M. S. (2014). Bayesian Modeling and Forecasting of 24-Hour High-Frequency Volatility. *Journal of the American Statistical Association*, **109**(508), 1368–1384.

Szerszen, P. J. (2009). Bayesian Analysis of Stochastic Volatility Models with Lévy Jumps: Application to Risk Analysis. *Working Paper*.

Takahashi, M., Omori, Y., and Watanabe, T. (2009). Estimating stochastic volatility models using daily returns and realized volatility simultaneously. *Computational Statistics and Data Analysis*, **53**, 2404–2426.

Wilhelmsson, A. (2013). Density Forecasting with Time-Varying Higher Moments: A model Confidence Set Approach. *Journal of Forecasting*, **31**(September), 19–31.

Yun, J. (2014). Out-of-sample density forecasts with affine jump diffusion models. *Journal of Banking & Finance*, **47**, 74–87.

A Simulation Study

We test the ability of the approximate maximum likelihood method to estimate the parameters of affine and non-affine specifications. This procedure extends simulation results in Bates (2006). We focus on a sample size of 4000 daily returns and simulate processes with 100 intra-daily time steps with an Euler discretization as in Eraker *et al.* (2003). We provide results for the standard stochastic volatility specification and an extension with state-depended jump probabilities.

Tables 18 and 19 report results for a small Monte Carol study with 100 random sample paths. The results indicate that the maximum likelihood method of Bates (2006) very accurately identifies the parameters of the stochastic variance and jump specifications. The local approximation to non-affine specifications leads to a minor loss in the precision of estimated parameters but the estimation methodology is still able to identify the parameters accurately.

[Table 15 about here.]

[Table 16 about here.]

B Implications for Value at Risk

In this section, we provide out-of-sample tests using a VaR-based loss function. Our aim is to understand the role of complex models for a standard application in financial risk management. To this end, we base out-of-sample tests on the asymmetric VaR loss function of González-Rivera *et al.* (2004). This function penalizes return observations below VaR more than return observations that are above VaR. For the details see Section 4.3.

[Table 17 about here.]

We first present MCS estimations for a VaR loss function with a significance level $\alpha = 1\%$, as this is the most common level used for financial applications. We report estimation results for all model classes in Table 20. The 25% model confidence set consists of five models, DGARCH models with fat-tailed error terms (GJR-t, MF-GJR-t, GJR-N-J) and two simple stochastic volatility models (SV-A, SV-G). All remaining models are contained in the 10% model confidence set. These findings can be interpreted as follows. First, complex continuous-time models do not provide any improvement over simpler DGARCH specifications as far as VaR estimations are concerned. Interestingly, simple DGARCH specifications (in particular GJR-t) outperform all jump-augmented continuous-time specifications. Secondly, the best-performing continuous-time models are SV-A and SV-G, a finding that supports earlier evidence in favor of these two specifications.

In order to test these results for robustness, we rerun the analysis for two further significance levels $\alpha = 0.5\%$ and $\alpha = 2\%$ (unreported).[26] Interestingly, the smaller the significance level, the more significant is the outperformance of the DGARCH specifications. For $\alpha = 0.5\%$, the 25% model confidence set consists of GJR-t, MF-GJR-t and GJR-N-J and the only additional model in the 10% model confidence set is GJR-N. Therefore for small significance levels, we find that simple DGARCH models significantly outperform continuous-time models. For higher α-levels, the choice of model is less important, as for $\alpha = 2\%$, we find that all but two models (MF-GJR-N-J, MF-SV-A) are included in the 25% MCS. Overall, this finding suggests that while continuous-time models provide significant improvements when the loss function takes into account the whole density (such as the predictive log-likelihood or the unweighted CRPS statistic), simple DGARCH models with fat error terms are superior for applications that focus on the performance of the left tail only.

[26] Detailed results for these tests are available upon request.

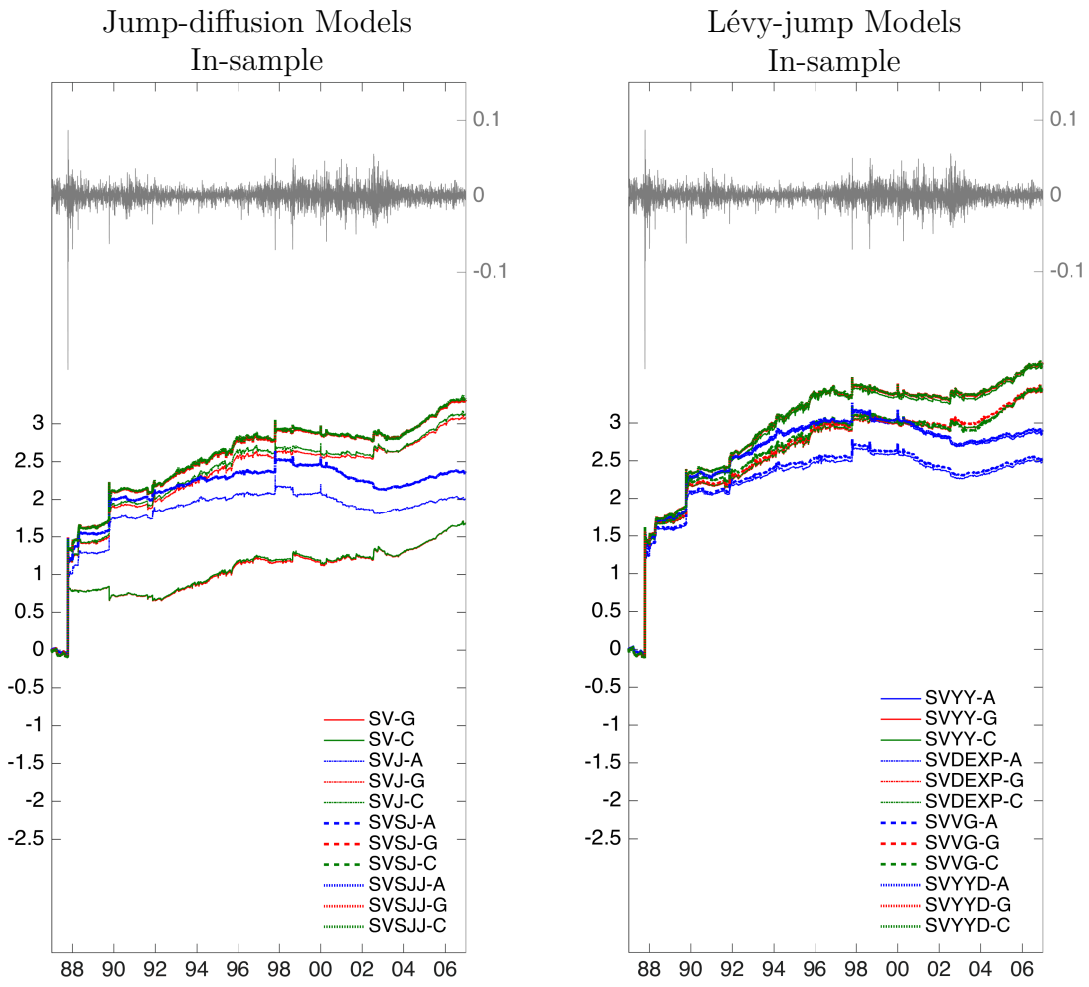


Figure 1: In-sample Sequential Likelihood Ratios.

These graphs show S&P 500 index returns in the upper part of the graphs and sequential likelihood ratios in the lower part of the graphs for the in-sample time period from January 2, 1987 to December 30, 2006. The left graph shows results for single factor jump-diffusion models and the right graph for the Lévy-jump models, respectively. Sequential likelihoods are calculated as a byproduct of the filtering procedure proposed by Bates (2006). All sequential likelihood ratios are calculated relative to the benchmark model SV-A.

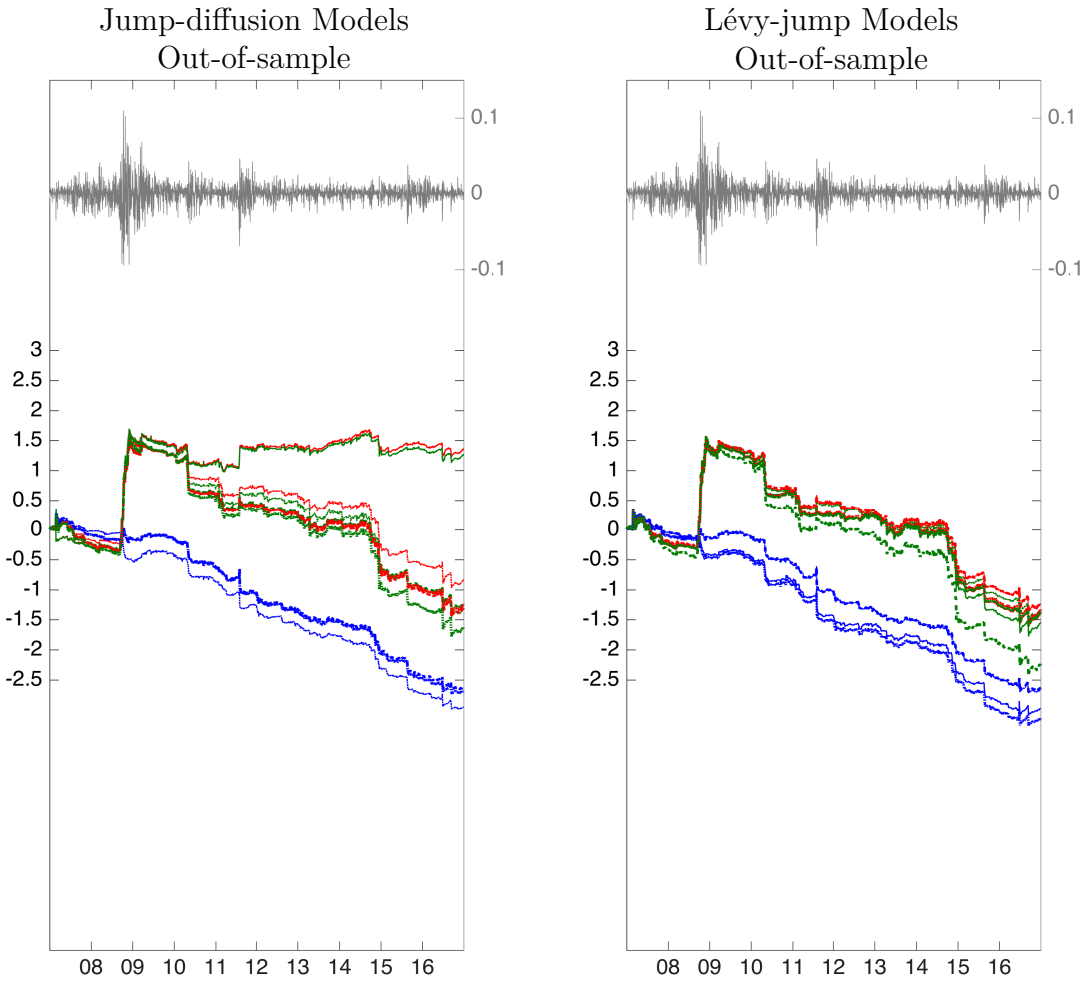


Figure 2: Out-of-sample Sequential Likelihood Ratios.

These graphs show S&P 500 index returns in the upper part of the graphs and sequential likelihood ratios in the lower part of the graphs for the out-of-sample time period from January 3, 2007 to December 30, 2016. The left graph shows results for the estimated single factor jump-diffusion models and the right graph for the Lévy-jump models, respectively. Sequential likelihoods are calculated as a byproduct of the employed estimation procedure proposed by Bates (2006). All sequential likelihood ratios are calculated relative to the benchmark model SV-A.

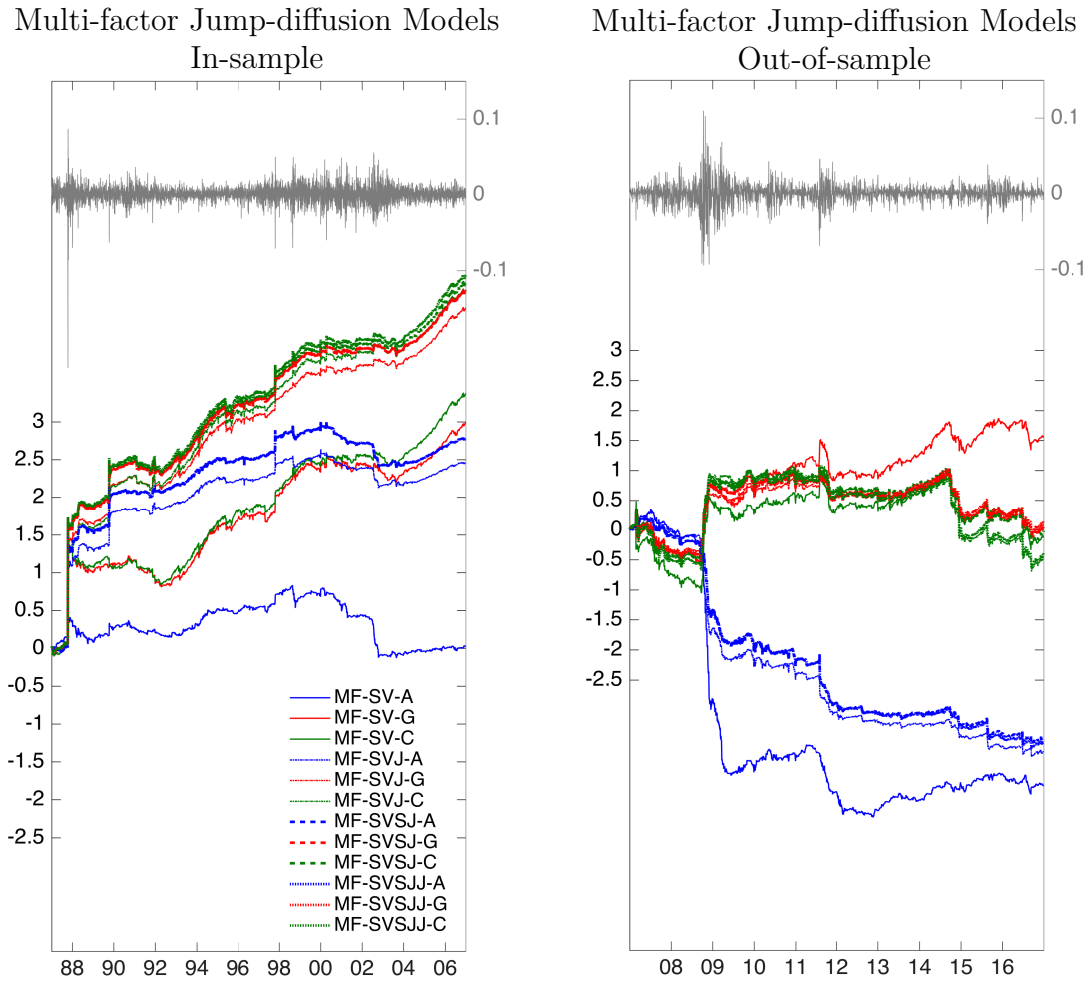
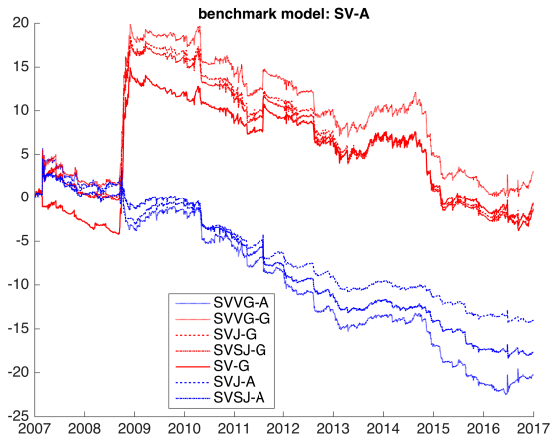


Figure 3: In-sample and Out-of-sample Sequential Likelihood Ratios for multi-factor jump-diffusion models.

The graphs show S&P 500 index returns in the upper part of the graphs and sequential likelihood ratios for the estimated multi-factor jump-diffusion models in the lower part of the graphs. The left graph shows the sequential likelihood ratios for the in-sample time period from January 2, 1987 until December 29, 2006 and the right graph for the Lévy-jump models for the out-of-sample time period January 3, 2007 until December 30, 2016.

(a) Model class: one-factor JD models



(b) Model class: two-factor JD models

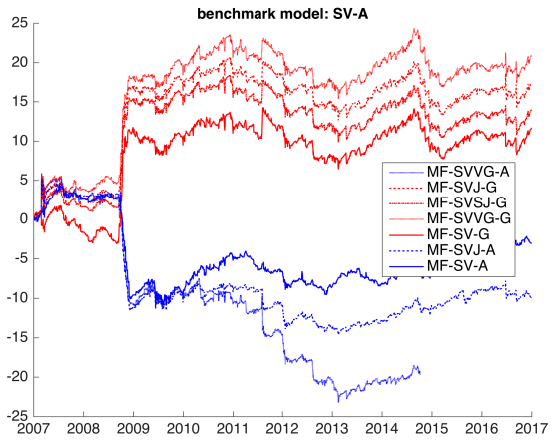
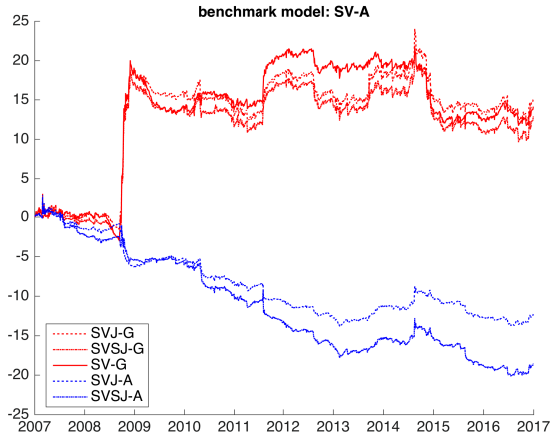


Figure 4: Out-of-sample log Bayes Factors: Updating Frequency

These graphs show sequential log Bayes factors for the out-of-sample time period from January 2, 2007 to December 30, 2016. The left graph shows results for single factor jump-diffusion models and the right graph for the two factor jump-diffusion models, respectively. Sequential log Bayes factors are calculated according to Equation (11). All sequential log Bayes factors are calculated relative to the benchmark model SV-A.

(c) Model class: 1FJD (RV)



(d) Model class: 1FJD (VIX)

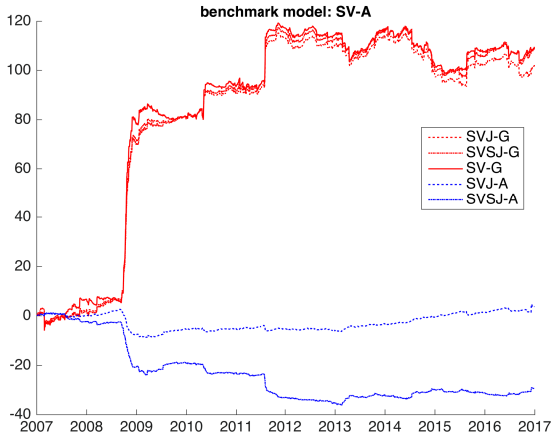
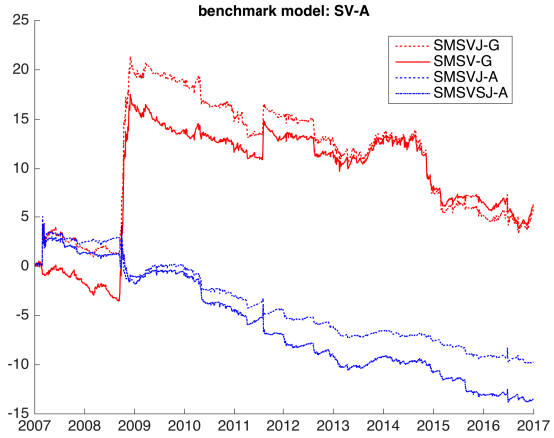


Figure 5: Out-of-sample log Bayes Factors: Additional Data

These graphs show sequential log Bayes factors for the out-of-sample time period from January 2, 2007 to December 30, 2016. The left graph shows results for single factor jump-diffusion models with the information set augmented by a realized variance estimator and the right graph for single factor jump-diffusion models with the information set augmented by the VIX index, respectively. Sequential log Bayes factors are calculated according to Equation (11). All sequential log Bayes factors are calculated relative to the benchmark model SV-A.

(a) Model class: one-factor JD models



(b) Model class: two-factor JD models

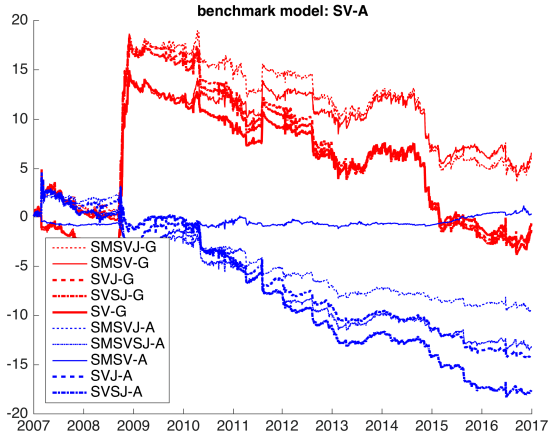


Figure 6: Out-of-sample log Bayes Factors: Time Varying Mean

These graphs show sequential log Bayes factors for the out-of-sample time period from January 2, 2007 to December 30, 2016. The left graph shows results for single factor jump-diffusion models and the right graph for the two factor jump-diffusion models, respectively. Sequential log Bayes factors are calculated according to Equation (11). In both cases the return equation was augmented by a time varying mean. All sequential log Bayes factors are calculated relative to the benchmark model SV-A.

Table 4: One-Factor Continuous-Time Models

This table provides an overview of the one-factor continuous-time models used in this paper. Panel A lists all jump-diffusion models, whereas Panel B provides specifications built from the CGMY process of Carr *et al.* (2002). Column 1 provides the model number, column 2 the acronym used throughout the paper and column 3 provides a short description of the main model features.

Number	Model	Features

Panel A: One-factor jump diffusion models. Models 1 to 12 are nested in the following SDEs (where λ_t is the intensity of N):

$$\begin{aligned} ds_t &= \left(\mu_c - \tfrac{1}{2}v_t - \lambda_t \bar{k}\right) dt + \rho_v \sqrt{v_t} dW_t^v + \sqrt{1-\rho_v^2}\sqrt{v_t}\, dW_t^s + \xi_t dN_t \\ dv_t &= \kappa_v \left(\theta_v - v_t\right) dt + \sigma_v v_t^{\gamma} dW_t^v. \end{aligned}$$

1	SV-A	Stochastic volatility model of Heston (1993), $\lambda_t = 0$ for all t, $\gamma = \frac{1}{2}$
2	SV-G	Continuous-time GARCH model with $\lambda_t = 0$ for all t, $\gamma = 1$
3	SV-C	CEV stochastic volatility model with with $\lambda_t = 0$ for all t, $\gamma \in [0.5, 1.5]$
4	SVJ-A	As model 1 with jump intensity $\lambda_t = \lambda_c$, normally distributed jump size ξ_t
5	SVJ-G	As model 2 with $\lambda_t = \lambda_c$, normally distributed jump size ξ_t
6	SVJ-C	As model 3 with $\lambda_t = \lambda_c$, normally distributed jump size ξ_t
7	SVSJ-A	As model 1 with $\lambda_t = \lambda_v v_t$, normally distributed jump size ξ_t
8	SVSJ-G	As model 2 with $\lambda_t = \lambda_v v_t$, normally distributed jump size ξ_t
9	SVSJ-C	As model 3 with $\lambda_t = \lambda_v v_t$, normally distributed jump size ξ_t
10	SVSJJ-A	As model 1 with $\lambda_t = \lambda_c + \lambda_v v_t$ normally distributed jump size ξ_t
11	SVSJJ-G	As model 2 with $\lambda_t = \lambda_c + \lambda_v v_t$, normally distributed jump size ξ_t
12	SVSJJ-C	As model 3 with $\lambda_t = \lambda_c + \lambda_v v_t$, normally distributed jump size ξ_t

Panel B: One-factor Levy-jump models. Models 13 to 24 are described by the following SDEs:

$$\begin{aligned} ds_t &= \left(\mu_c - \tfrac{1}{2}v_t\right) dt + \rho_v \sqrt{v_t} dW_t^v + \sqrt{1-\rho_v^2}\sqrt{v_t}\, dL_t \\ dv_t &= \kappa_v \left(\theta_v - v_t\right) dt + \sigma_v v_t^{\gamma} dW_t^v. \end{aligned}$$

13	SVYY-A	L_t driven by CGMY process of Carr *et al.* (2003), $\gamma = \frac{1}{2}$
14	SVYY-G	L_t driven by CGMY process of Carr *et al.* (2003), $\gamma = 1$
15	SVYY-C	L_t driven by CGMY process of Carr *et al.* (2003), $\gamma \in [0.5, 1.5]$
16	SVDEXP-A	L_t driven by double exponential jumps as in Kou (2002), $\gamma = \frac{1}{2}$
17	SVDEXP-G	L_t driven by double exponential jumps as in Kou (2002), $\gamma = 1$
18	SVDEXP-C	L_t driven by double exponential jumps as in Kou (2002), $\gamma \in [0.5, 1.5]$
19	SVVG-A	L_t driven by VG process of Madan and Seneta (1990), $\gamma = \frac{1}{2}$
20	SVVG-G	L_t driven by VG process of Madan and Seneta (1990), $\gamma = 1$
21	SVVG-C	L_t driven by VG process of Madan and Seneta (1990), $\gamma \in [0.5, 1.5]$
22	SVYYD-A	As model 13 with additional diffusive component in L_t
23	SVYYD-G	As model 14 with additional diffusive component in L_t
24	SVYYD-C	As model 15 with additional diffusive component in L_t

Table 5: Data Statistics

This table provides summary statistics for daily log returns of the S&P 500 index for the whole sample period from January 2, 1987 to December 30, 2016 as well as various sub-samples. In particular the sample in column 4 (with sample start 1987*) excludes the observation on October 16, 1987, a market crash with a log return of -22.9%.

Sample start	1987	1987	1987*	2007	2006	2010
Sample end	2016	2006	2006	2014	2009	2016
Observations	7561	5043	5042	2518	757	1761
Mean	0.0003	0.0003	0.0004	0.0002	-0.0003	0.0004
Standard deviation	0.0116	0.0108	0.0103	0.0132	0.0189	0.0098
Skewness	-1.2686	-2.0918	-0.2124	-0.3265	-0.1737	-0.4381
Kurtosis	30.7534	48.3125	8.9613	12.9216	9.0731	7.2018
Percentile 0.5%	-0.0408	-0.0330	-0.0321	-0.0515	-0.0764	-0.0344
Percentile 1%	-0.0313	-0.0273	-0.0272	-0.0405	-0.0588	-0.0288
Percentile 2%	-0.0250	-0.0226	-0.0226	-0.0304	-0.0479	-0.0231
Percentile 5%	-0.0173	-0.0161	-0.0161	-0.0206	-0.0300	-0.0160
Percentile 50%	0.0006	0.0005	0.0005	0.0006	0.0009	0.0005
Percentile 95%	0.0166	0.0158	0.0158	0.0178	0.0263	0.0151
Percentile 98%	0.0237	0.0223	0.0223	0.0283	0.0400	0.0207
Percentile 99%	0.0308	0.0276	0.0276	0.0375	0.0526	0.0251
Percentile 99.5%	0.0383	0.0347	0.0347	0.0430	0.0657	0.0319

Table 6: Model Confidence Set p-Values and Model Ranking Full Out-of-sample Period Using Predictive Likelihood

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using predictive likelihood as the ranking criteria. For details regarding notation, see Section 4.2 and Section 4.3. The first column indicates the number of the iterative elimination step for models running from $i = 1$ to total number of models ($m_0 = 14$). The second column shows the p-values for the hypotheses $H_{0,\mathcal{M}_i}$ and the third column presents the MCS p-value $\hat{p}_{e_{\mathcal{M}_i}}$ for the model that is removed in the respective elimination step. The fourth column shows the model eliminated in each iterative step by the elimination rule and thereby presents the model ranking according the MCS criteria, with the worst model ranked at the top and the best model at the bottom of the table, respectively. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.0036	0.0036	SVYYD-A
$e_{\mathcal{M}_2}$	0.0031	0.0036	SVYY-A
$e_{\mathcal{M}_3}$	0.0026	0.0036	SVDEXP-A
$e_{\mathcal{M}_4}$	0.0023	0.0036	SVSJ-A
$e_{\mathcal{M}_5}$	0.0021	0.0036	SVVG-A
$e_{\mathcal{M}_6}$	0.0017	0.0036	SVJ-A
$e_{\mathcal{M}_7}$	0.0009	0.0036	SVSJ-G
$e_{\mathcal{M}_8}$	0.0013	0.0036	SVYY-G
$e_{\mathcal{M}_9}$	0.0015	0.0036	SVDEXP-G
$e_{\mathcal{M}_{10}}$	0.0026	0.0036	SVVG-G
$e_{\mathcal{M}_{11}}$	0.0062	0.0062	SVYYD-G
$e_{\mathcal{M}_{12}}$	0.0335	0.0335	SVJ-G
$e_{\mathcal{M}_{13}}$	0.2509	0.2509	SV-A
$e_{\mathcal{M}_{14}}$	1.0000	1.0000	SV-G

Table 7: Model Confidence Set p-Values and Model Ranking First Part and Second Part of Out-of-sample Using Predictive Likelihood

This table shows model confidence set results for the first part of the out-of-sample period January 3, 2007 to December 31, 2009 in the upper panel and the second part of the out-of-sample period January 4, 2010 to December 30, 2016 in the lower panel using predictive likelihood as the ranking criteria. For details regarding notation, see Section 4.2 and Section 4.3. The first column indicates the number of the iterative elimination step for models running from $i = 1$ to total number of models ($m_0 = 14$). The second column shows the p-values for the hypotheses $H_{0,\mathcal{M}_i}$ and the third column presents the MCS p-value $\hat{p}_{e_{\mathcal{M}_i}}$ for the model that is removed in the respective elimination step. The fourth column shows the model eliminated in each iterative step by the elimination rule and thereby presents the model ranking according the MCS criteria, with the worst model ranked at the top and the best model at the bottom of the table, respectively. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Panel A: January 2007 to December 2009

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.1504	0.1504	SVYYD-A
$e_{\mathcal{M}_2}$	0.1499	0.1504	SVYY-A
$e_{\mathcal{M}_3}$	0.1504	0.1504	SVSJ-A
$e_{\mathcal{M}_4}$	0.1497	0.1504	SVDEXP-A
$e_{\mathcal{M}_5}$	0.1499	0.1504	SVVG-A
$e_{\mathcal{M}_6}$	0.1535	0.1535	SVJ-A
$e_{\mathcal{M}_7}$	0.1937	0.1937	SV-A
$e_{\mathcal{M}_8}$	0.5746	0.5746	SVSJ-G
$e_{\mathcal{M}_9}$	0.5490	0.5746	SVYY-G
$e_{\mathcal{M}_{10}}$	0.7044	0.7044	SVYYD-G
$e_{\mathcal{M}_{11}}$	0.7812	0.7812	SVDEXP-G
$e_{\mathcal{M}_{12}}$	0.7846	0.7846	SVVG-G
$e_{\mathcal{M}_{13}}$	0.9958	0.9958	SVJ-G
$e_{\mathcal{M}_{14}}$	1.0000	1.0000	SV-G

Panel B: January 2010 to December 2016

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.0006	0.0006	SVYYD-A
$e_{\mathcal{M}_2}$	0.0005	0.0006	SVYY-G
$e_{\mathcal{M}_3}$	0.0006	0.0006	SVDEXP-A
$e_{\mathcal{M}_4}$	0.0005	0.0006	SVSJ-G
$e_{\mathcal{M}_5}$	0.0009	0.0009	SVVG-A
$e_{\mathcal{M}_6}$	0.0007	0.0009	SVVG-G
$e_{\mathcal{M}_7}$	0.0013	0.0013	SVJ-A
$e_{\mathcal{M}_8}$	0.0008	0.0013	SVDEXP-G
$e_{\mathcal{M}_9}$	0.0025	0.0025	SVSJ-A
$e_{\mathcal{M}_{10}}$	0.0019	0.0025	SVYYD-G
$e_{\mathcal{M}_{11}}$	0.0071	0.0071	SVYY-A
$e_{\mathcal{M}_{12}}$	0.0007	0.0071	SVJ-G
$e_{\mathcal{M}_{13}}$	0.8981	0.8981	SV-G
$e_{\mathcal{M}_{14}}$	1.0000	1.0000	SV-A

Table 8: Gneiting-Ranjan Tests
Full Out-of-sample Dataset (No weighting).

This table reports the Gneiting and Ranjan (2011) test statistics $t_n = \sqrt{n}\left(\overline{CRPS}_w^f - \overline{CRPS}_w^g\right)\hat{\sigma}_n^{-1}$ for several model pairs with CRPS denoting continuous ranked probability score and f and g denoting forecasting densities of the models to be tested against each other. The test statistic follows asymptotically a standard normal distribution. For details of calculation, see Section 4.3. The models in rows refer to forecasting density f and the models in columns to forecasting density g, respectively. A positive statistic therefore indicates that the model in the row is out-performed by the model in the column and vice versa.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
SV-A (1)	–	2.44	-4.16	1.48	-3.76	1.31	-3.28	1.51	-3.73	1.34
SV-G (2)	-2.44	–	-3.46	-3.51	-3.25	-3.10	-3.14	-2.80	-3.26	-3.10
SVJ-A (3)	4.16	3.46	–	2.95	-1.45	2.50	-0.27	2.63	-1.55	2.51
SVJ-G (4)	-1.48	3.51	-2.95	–	-3.04	-2.99	-2.92	-2.03	-3.06	-2.92
SVYY-A (5)	4.44	3.60	1.39	3.11	–	2.73	2.50	2.83	-0.85	2.74
SVYY-G (6)	-0.91	3.92	-2.52	3.13	-2.73	–	-2.58	2.80	-2.75	2.14
SVVG-A (7)	3.97	3.47	0.03	2.95	-2.50	2.58	–	2.70	-3.04	2.59
SVVG-G (8)	-1.14	3.66	-2.68	2.08	-2.83	-2.80	-2.70	–	-2.86	-2.54
SVYYD-A (9)	4.43	3.62	1.54	3.14	0.85	2.75	3.04	2.86	–	2.76
SVYYD-G (10)	-0.95	3.91	-2.55	3.01	-2.74	-2.14	-2.59	2.54	-2.76	–

Table 9: Gneiting-Ranjan Tests
Full Out-of-sample Dataset Dataset (left tail weighting).

This table reports the Gneiting and Ranjan (2011) test statistics $t_n = \sqrt{n}\left(\overline{CRPS}_w^f - \overline{CRPS}_w^g\right)\hat{\sigma}_n^{-1}$ for several model pairs with CRPS denoting continuous ranked probability score with weight function $w(\alpha) = (1-\alpha)^2$ and f and g denoting forecasting densities of the models to be tested against each other. The test statistic follows asymptotically a standard normal distribution. For details of calculation see Section 4.3. The models in rows refer to forecasting density f and the models in columns to forecasting density g, respectively. A positive statistic therefore indicates that the model in the row is out-performed by the model in the column and vice versa.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
SV-A (1)	–	1.20	-2.10	0.96	-2.03	0.78	-1.77	0.90	-2.03	0.80
SV-G (2)	-1.20	–	-1.72	-0.99	-1.59	-1.07	-1.50	-0.84	-1.60	-1.04
SVJ-A (3)	2.10	1.72	–	1.68	-0.77	1.43	0.13	1.51	-0.81	1.44
SVJ-G (4)	-0.96	0.99	-1.68	–	-1.63	-1.47	-1.53	-0.96	-1.64	-1.43
SVYY-A (5)	2.19	1.76	0.62	1.72	–	1.53	1.74	1.60	-0.33	1.54
SVYY-G (6)	-0.71	1.35	-1.53	1.54	-1.53	–	-1.41	1.68	-1.54	1.25
SVVG-A (7)	1.99	1.69	-0.28	1.63	-1.74	1.41	–	1.49	-1.94	1.42
SVVG-G (8)	-0.85	1.14	-1.63	0.98	-1.60	-1.68	-1.49	–	-1.60	-1.52
SVYYD-A (9)	2.21	1.78	0.71	1.73	0.33	1.54	1.94	1.60	–	1.55
SVYYD-G (10)	-0.73	1.33	-1.54	1.48	-1.54	-1.25	-1.42	1.52	-1.55	–

Table 10: Model Confidence Set p-Values and Model Ranking Full Out-of-sample Period Using CRPS

This table shows model confidence set results for the full out-of-sample period from January 3, 2007 to December 30, 2016 using continuous ranked probability score (CRPS) as the ranking criteria. For details of notation and calculation see Sections 4.2 and 4.3. The first column provides model specifications, the second column provides results for the non-weighted CRPS statistic. Columns 3 to 6 refer to the results for the weighted CRPS statistics. The weighting scheme "Center" applies more weight to the center of the predictive density when calculating CRPS and the weighting schemes "Tails", "Right Tail", and "Left Tail" work accordingly. For a given significance level α models for which $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ are included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$. We use $*$ ($**$) to indicate that the model belongs to the 10% (25%) MCS.

Model Name	No Weight	Center	Tails	Right Tail	Left Tail
SV-A	0.0294	0.0188	0.0544	0.0084	0.4101**
SV-G	1.0000**	1.0000**	1.0000**	1.0000**	1.0000**
SVJ-A	0.0294	0.0188	0.0544	0.0084	0.3069**
SVJ-G	0.0294	0.0188	0.0544	0.0084	0.4101**
SVSJ-A	0.0294	0.0188	0.0544	0.0084	0.2840**
SVSJ-G	0.0294	0.0188	0.0544	0.0084	0.4101**
SVYY-A	0.0294	0.0188	0.0544	0.0084	0.2840**
SVYY-G	0.0294	0.0188	0.0544	0.0084	0.4101**
SVDEXP-A	0.0294	0.0188	0.0544	0.0084	0.2840**
SVDEXP-G	0.0294	0.0188	0.0544	0.0084	0.4101**
SVVG-A	0.0294	0.0188	0.0544	0.0084	0.2840**
SVVG-G	0.0294	0.0188	0.0544	0.0084	0.4101**
SVYYD-A	0.0294	0.0188	0.0544	0.0084	0.2840**
SVYYD-G	0.0294	0.0188	0.0544	0.0084	0.4101**

Table 11: Model Confidence Set p-Values and Model Ranking First Part and Second Part of Out-of-sample Using CRPS

This table provides model confidence set results for the first part of out-of-sample period from January 3, 2007 to December 31, 2009 in panel A and the second part of the out-of-sample period January 4, 2010 to December 30, 2016 in Panel B. The loss function is given by the continuous ranked probability score (CRPS). For details of notation and calculation see Sections 4.2 and 4.3. The first column provides model specifications, the second column provides results for the non-weighted CRPS statistic. Columns 3 to 6 refer to the results for the weighted CRPS statistics. The weighting scheme "Center" applies more weight to the center of the predictive density when calculating CRPS and the weighting schemes "Tails", "Right Tail", and "Left Tail" work accordingly. For a given significance level α models for which $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ are included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$. We use $*$ ($**$) to indicate that the model belongs to the 10% (25%) MCS.

Panel A: January 2007 to December 2009

Model Name	No Weight	Center	Tails	Right Tail	Left Tail
SV-A	0.2565**	0.2746**	0.1739*	0.0868	0.3322**
SV-G	1.0000**	1.0000**	1.0000**	1.0000**	0.9565**
SVJ-A	0.1393*	0.1437*	0.1489*	0.0629	0.3322**
SVJ-G	0.4573**	0.4730**	0.2869**	0.1175*	1.0000**
SVSJ-A	0.1675*	0.1707*	0.1489*	0.0629	0.3322**
SVSJ-G	0.4400**	0.4730**	0.2869**	0.1175*	0.8719**
SVYY-A	0.1284*	0.1437*	0.1370*	0.0629	0.3293**
SVYY-G	0.2565**	0.4730**	0.2285*	0.0974	0.7440**
SVDEXP-A	0.1474*	0.1464*	0.1489*	0.0629	0.3322**
SVDEXP-G	0.4573**	0.4730**	0.2869**	0.1175*	0.9565**
SVVG-A	0.1529*	0.1437*	0.1489*	0.0629	0.3394**
SVVG-G	0.4573**	0.4730**	0.2869**	0.0974	0.9003**
SVYYD-A	0.1300*	0.1437*	0.1370*	0.0629	0.3322**
SVYYD-G	0.2565**	0.4730**	0.2285*	0.0868	0.8047**

Panel B: January 2010 to December 2016

Model Name	No Weight	Center	Tails	Right Tail	Left Tail
SV-A	0.2085*	0.0581	0.6423**	0.1747*	0.8857**
SV-G	1.0000**	1.0000**	1.0000**	1.0000**	1.0000**
SVJ-A	0.0151	0.0019	0.0443	0.0154	0.2436*
SVJ-G	0.0151	0.0025	0.0704	0.0200	0.3826**
SVSJ-A	0.0151	0.0024	0.0704	0.0200	0.2387*
SVSJ-G	0.0151	0.0025	0.0535	0.0154	0.2436*
SVYY-A	0.0151	0.0022	0.0326	0.0154	0.2436*
SVYY-G	0.0151	0.0025	0.0352	0.0154	0.2436*
SVDEXP-A	0.0151	0.0023	0.0352	0.0154	0.2387*
SVDEXP-G	0.0151	0.0027	0.0443	0.0154	0.2436*
SVVG-A	0.0151	0.0022	0.0352	0.0154	0.2387*
SVVG-G	0.0151	0.0025	0.0443	0.0154	0.3092**
SVYYD-A	0.0151	0.0019	0.0298	0.0154	0.2387*
SVYYD-G	0.0151	0.0025	0.0352	0.0154	0.2501**

Table 12: Likelihood Ratio Tests (Sub-Sample Analysis).
H_0: $\mu = \rho = 0$ and $\sigma = 1$

This table reports the likelihood ratios (LR) for the likelihood ratio test proposed in Berkowitz (2001) as described in Section (6.4). Based on the estimated optimal parameters sets for time period 1987 to end of 2006, LRs are calculated for the full out-of-sample period from start of 2007 until end of 2016 and for each of the out-of-sample years separately. The LR is calculated as $LR = -2[\mathcal{L}(0,1,0) - \mathcal{L}(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})]$, which serve as test statistics for the null hypothesis H0: ($\mu = 0$, $\sigma^2 = 1$, $\rho = 0$), jointly testing the probability integral transforms for independence and mean and variance equal to $(0, 1)$. The test statistic is distributed $\chi^2(3)$ with critical values given by: 99% level – $\chi^2(3) = 11.34$ (***), 95% level: $\chi^2(3) = 9.210$ (**), and 90% level: $\chi^2(3) = 6.635$ (*).

Model	All	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
SV-A	21.13***	5.36	34.61***	5.11	3.12	7.39*	0.79	6.70*	1.01	0.89	5.04
SV-G	24.07***	7.40*	13.65***	3.01	5.93	7.29*	0.27	6.85*	0.77	1.99	3.89
SVJ-A	20.91***	6.45	33.17***	2.38	2.48	6.60	0.91	6.09	0.45	0.96	5.31
SVJ-G	22.46***	8.08*	13.89***	2.05	4.33	6.03	0.14	6.25	1.24	2.06	4.02
SVSJ-A	19.66***	6.39	26.01***	2.22	2.38	6.89*	0.75	6.13	0.42	0.99	4.91
SVSJ-G	21.67***	8.13*	12.65***	1.65	4.07	6.24	0.16	6.47	1.50	1.94	3.94
SVYY-A	23.49***	7.09*	30.83***	2.66	2.80	8.99*	0.78	5.94	0.45	1.04	4.92
SVYY-G	24.72***	9.00*	13.73***	1.76	4.56	7.84*	0.15	6.37	1.75	1.94	4.14
SVDEXP-A	20.08***	6.51	26.43***	2.17	2.36	7.27*	0.80	6.09	0.43	1.02	4.84
SVDEXP-G	22.34***	8.26*	12.80***	1.68	4.11	6.51	0.16	6.51	1.55	1.96	3.94
SVVG-A	20.06***	6.56	26.09***	2.14	2.39	7.29*	0.83	6.19	0.44	0.98	4.88
SVVG-G	22.22***	8.29*	12.86***	1.68	4.09	6.57	0.16	6.55	1.51	1.89	4.04
SVYYD-A	23.34***	7.05*	30.76***	2.54	2.75	8.97*	0.77	5.93	0.44	1.04	4.88
SVYYD-G	24.67***	8.96*	13.69***	1.78	4.60	7.78*	0.15	6.38	1.74	1.94	4.12

Table 13: Model Confidence Set p-Values and Model Ranking Multi-factor Models for Full Out-of-sample Period using Predictive Likelihood

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using predictive likelihood as the ranking criteria. For details of notation and calculation see Section 4.2 and Section 4.3. Multi-factor models are tested against SV-A, SV-G, SVSJJ-A, SVSJJ-G, SVYYD-A, and SVYYD-G. The first column indicates the number of the iterative elimination step for models running from $i = 1$ to total number of models ($m_0 = 17$). The second column shows the p-values for the hypotheses $H_{0,\mathcal{M}_i}$ and the third column presents the MCS p-value $\hat{p}_{e_{\mathcal{M}_i}}$ for the model that is removed in the respective elimination step. The fourth column shows the model eliminated in each iterative step by the elimination rule and thereby presents the model ranking according to the MCS criteria, with the worst model ranked at the top and the best model at the bottom of the table respectively. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.0147	0.0147	MF-SVSJJ-A
$e_{\mathcal{M}_2}$	0.0170	0.0170	SVYYD-A
$e_{\mathcal{M}_3}$	0.0172	0.0172	MF-SVSJ-A
$e_{\mathcal{M}_4}$	0.0221	0.0221	SVSJ-A
$e_{\mathcal{M}_5}$	0.0234	0.0234	MF-SVJ-A
$e_{\mathcal{M}_6}$	0.0333	0.0333	MF-SV-A
$e_{\mathcal{M}_7}$	0.1263	0.1263	SVSJ-G
$e_{\mathcal{M}_8}$	0.1372	0.1372	SVYYD-G
$e_{\mathcal{M}_9}$	0.1416	0.1416	MF-SVSJJ-C
$e_{\mathcal{M}_{10}}$	0.1511	0.1511	MF-SVJ-C
$e_{\mathcal{M}_{11}}$	0.1630	0.1630	MF-SVSJ-C
$e_{\mathcal{M}_{12}}$	0.1623	0.1630	MF-SVSJ-G
$e_{\mathcal{M}_{13}}$	0.1602	0.1630	MF-SV-C
$e_{\mathcal{M}_{14}}$	0.3485	0.3485	MF-SVJ-G
$e_{\mathcal{M}_{15}}$	0.4291	0.4291	MF-SVSJJ-G
$e_{\mathcal{M}_{16}}$	0.3884	0.4291	SV-A
$e_{\mathcal{M}_{17}}$	0.8147	0.8147	SV-G
$e_{\mathcal{M}_{18}}$	1.0000	1.0000	MF-SV-G

Table 14: Model Confidence Set p-Values and Model Ranking Multi-factor Models for Full Out-of-sample Period Using CRPS

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using continuous ranked probability score (CRPS) as the ranking criteria. For details of notation and calculation see Sections 4.2 and Section 4.3. Multi-factor models are tested against SV-A, SV-G, SVSJJ-A, SVSJJ-G, SVYYD-A, and SVYYD-G. First column gives models tested listed from least complex model at the top to most complex model at the bottom of table. Column 2 gives results of the non-weighted CRPS version. Columns 3 to 6 refer to the results of the weighted CRPS versions. The weighting scheme "Center" puts more weight on the center of the predictive density when calculating CRPS and the weighting schemes "Tails", "Right Tail", and "Left Tail" work accordingly. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$. One $*$ indicates the model belongs to the 10% MCS and two $**$ indicate model belongs to the 25% MCS.

Model Name	No Weight	Center	Tails	Right Tail	Left Tail
MF-SV-A	0.0556	0.1852*	0.0575	0.0541	0.1817*
MF-SV-G	0.4536**	0.5236**	0.4715**	0.5948**	0.4798**
MF-SV-C	0.4536**	0.5236**	0.4715**	1.0000**	0.4543**
MF-SVJ-A	0.0307	0.0167	0.0645	0.0193	0.3986**
MF-SVJ-G	0.3910**	0.3488**	0.4276**	0.2355*	0.4798**
MF-SVJ-C	0.4536**	0.4290**	0.4715**	0.4356**	0.4798**
MF-SVSJ-A	0.1716*	0.1084*	0.2082*	0.0317	0.2574**
MF-SVSJ-G	0.3910**	0.4290**	0.4715**	0.3092**	0.4798**
MF-SVSJ-C	0.4536**	0.5236**	0.4715**	0.3814**	0.4798**
MF-SVSJJ-A	0.0352	0.0286	0.0940	0.0231	0.1979*
MF-SVSJJ-G	0.4536**	0.5075**	0.4715**	0.3413**	0.4798**
MF-SVSJJ-C	0.4536**	0.5185**	0.4715**	0.3814**	0.4798**
SV-A	0.2814**	0.2938**	0.3304**	0.2023*	0.4798**
SV-G	1.0000**	1.0000**	1.0000**	0.5948**	1.0000**
SVSJ-A	0.1516*	0.0200	0.2786**	0.0193	0.4798**
SVSJ-G	0.3910**	0.4290**	0.4276**	0.0880	0.4798**
SVYYD-A	0.0279	0.0148	0.0575	0.0115	0.4531**
SVYYD-G	0.3130**	0.3488**	0.3479**	0.0541	0.4798**

Table 15: In-sample parameter estimation results (Discrete-time GARCH Models).

This table reports the parameter estimation results for discrete-time GARCH models. The estimation period is from January 2, 1987 to December 29, 2006. The estimation is performed using maximum likelihood method. For each parameter, we report the maximum likelihood estimates and the standard errors in parenthesis. Log-likelihood values for each model are given in the last row. For exact model definitions see Section (7.2).

	GJR-N	MF-GJR-N	GJR-N-J	MF-GJR-N-J	GJR-t	MF-GJR-t
μ	0.0318	0.0393	0.0285	0.0287	0.0467	0.0472
	(0.0116)	(0.0113)	(0.0124)	(0.0132)	(0.0106)	(0.0105)
$\bar{q}$	1.0918		1.0396		0.9571	
	(0.1093)		(0.1412)		(0.1896)	
α_h	0.0136	0.0005	0.0196	0.0147	0.0178	0.0001
	(0.0058)	(0.0194)	(0.0067)	(0.0094)	(0.0089)	(0.0269)
β_h	0.9040	0.6100	0.8943	0.9212	0.9124	0.7552
	(0.0041)	(0.0494)	(0.0075)	(0.0174)	(0.0076)	(0.0705)
γ_h	0.1308	0.2122	0.1211	0.0900	0.1110	0.1233
	(0.0067)	(0.0212)	(0.0112)	(0.0232)	(0.0132)	(0.0373)
q_q		1.0810		0.5825		1.4592
		(0.1773)		(0.0581)		(0.8374)
α_q		0.0250		0.0008		0.0332
		(0.0076)		(0.0155)		(0.0144)
β_q		0.9519		0.8714		0.9487
		(0.0046)		(0.3938)		(0.0076)
γ_q		0.0289		-0.0002		0.0266
		(0.0123)		(0.0288)		(0.0244)
λ_b			0.0114	0.0065		
			(0.0035)	(0.0024)		
μ_r			-1.6804	-2.6934		
			(0.8805)	(1.7811)		
σ_r			2.8360	3.6652		
			(0.2697)	(0.4840)		
η					6.8810	6.8291
					(0.5240)	(0.5641)
LL	16623	16663	16741	16748	16774	16785

Table 16: Model Confidence Set p-Values and Model Ranking Discrete-time GARCH Models for Full Out-of-sample Period using Predictive Likelihood

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using predictive likelihood as the ranking criteria. For details of notation and calculation see Section 4.2 and Section 4.3. Discrete-time models are tested against SV-A, SV-G, SVSJ-A, SVSJ-G, SVYYD-A, and SVYYD-G. The first column indicates the number of the iterative elimination steps for models running from $i = 1$ to total number of models ($m_0 = 12$). Second column shows the p-values for the hypotheses $H_{0,\mathcal{M}_i}$ and third column presents MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$ for the model that is going to be eliminated in the respective elimination step. Fourth column shows the model eliminated in each iterative step by the elimination rule and thereby presents the model ranking according the MCS criteria, with the worst model ranked at the top and the best model at the bottom of the table, respectively. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.0000	0.0000	MF-GJR-J
$e_{\mathcal{M}_2}$	0.0000	0.0000	MF-GJR-N
$e_{\mathcal{M}_3}$	0.0000	0.0000	GJR-N
$e_{\mathcal{M}_4}$	0.0003	0.0003	GJR-N-J
$e_{\mathcal{M}_5}$	0.0010	0.0010	SVYYD-A
$e_{\mathcal{M}_6}$	0.0005	0.0010	SVSJ-A
$e_{\mathcal{M}_7}$	0.0002	0.0010	MF-GJR-t
$e_{\mathcal{M}_8}$	0.0005	0.0010	GJR-t
$e_{\mathcal{M}_9}$	0.0043	0.0043	SVSJ-G
$e_{\mathcal{M}_{10}}$	0.0300	0.0300	SVYYD-G
$e_{\mathcal{M}_{11}}$	0.2833	0.2833	SV-A
$e_{\mathcal{M}_{12}}$	1.0000	1.0000	SV-G

Table 17: Model Confidence Set p-Values and Model Ranking Discrete-time Models for Full Out-of-sample Period Using CRPS

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using continuous ranked probability score (CRPS) as the ranking criteria. For details of notation and calculation see Sections 4.2 and Section 4.3. The first column provides model specifications, the second column provides results for the non-weighted CRPS statistic. Columns 3 to 6 refer to the results for the weighted CRPS statistics. The weighting scheme "Center" applies more weight to the center of the predictive density when calculating CRPS and the weighting schemes "Tails", "Right Tail", and "Left Tail" work accordingly. For a given significance level α models for which $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ are included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$. We use $*$ ($**$) to indicate that the model belongs to the 10% (25%) MCS.

Model Name	No Weight	Center	Tails	Right Tail	Left Tail
GJR-N	0.0030	0.0014	0.0145	0.0006	0.6467**
MF-GJR-N	0.0013	0.0005	0.0084	0.0006	0.2061*
GJR-N-J	0.0030	0.0017	0.0823	0.0016	0.4694**
MF-GJR-J	0.0030	0.0017	0.0084	0.0006	0.2865**
GJR-t	0.0030	0.0057	0.0145	0.0006	1.0000**
MF-GJR-t	0.0030	0.0021	0.0084	0.0006	0.8170**
SV-A	0.0030	0.0057	0.0084	0.0016	0.4694**
SV-G	1.0000**	1.0000**	1.0000**	1.0000**	0.9237**
SVSJ-A	0.0030	0.0017	0.0084	0.0006	0.2865**
SVSJ-G	0.0061	0.0057	0.0145	0.0016	0.8170**
SVYYD-A	0.0030	0.0014	0.0084	0.0006	0.2672**
SVYYD-G	0.0030	0.0057	0.0084	0.0016	0.6467**

Table 18: Simulation study: SV model.

This table reports the parameter estimation results from a Monte Carlo study where 100 sample paths with 4000 daily returns are simulated from the true model with parameters shown as *simulated*. The simulation is performed using an Euler discretization with 100 time steps per day. The average estimated parameter of these simulated paths are reported in line *estimated*. RMSE and standard errors (*std error*) are also reported.

Parameter	μ	κ	θ	σ_v	ρ_v	γ
simulated	0.050	3.500	0.025	0.400	-0.650	0.500
estimated	0.041	3.796	0.025	0.399	-0.663	
RMSE	0.026	0.681	0.004	0.026	0.040	
standard error	0.003	0.068	0.000	0.003	0.004	
simulated	0.050	3.500	0.025	2.530	-0.650	1.000
estimated	0.043	4.381	0.023	2.404	-0.683	
RMSE	0.033	0.988	0.003	0.178	0.051	
standard error	0.003	0.099	0.000	0.018	0.005	
simulated	0.050	3.500	0.025	1.006	-0.650	0.750
estimated	0.042	3.850	0.024	1.051	-0.678	0.761
RMSE	0.032	0.830	0.004	0.308	0.046	0.073
standard error	0.003	0.083	0.000	0.031	0.005	0.007
simulated	0.050	3.500	0.025	3.658	-0.650	1.100
estimated	0.044	4.780	0.022	3.252	-0.683	1.070
RMSE	0.033	1.139	0.003	1.086	0.053	0.082
standard error	0.003	0.114	0.000	0.109	0.005	0.008

Table 19: Simulation study: SV model.

This table reports the parameter estimation results from a Monte Carlo study where 100 sample paths with 4000 daily returns are simulated from the true model with parameters shown as *simulated*. The simulation is performed using an Euler discretization with 100 time steps per day. The average estimated parameter of these simulated paths are reported in line *estimated*. RMSE and standard errors (*std error*) are also reported.

Parameter	μ	κ	θ	σ_v	ρ_v	λ_c	λ_v	μ_s	σ_s	γ
sim	0.050	3.500	0.025	0.400	-0.650	0.500	30.000	-0.020	0.050	0.500
est	0.046	3.765	0.025	0.400	-0.657	0.546	34.957	-0.024	0.043	
RMSE	0.026	0.722	0.004	0.028	0.044	0.439	26.084	0.024	0.014	
std err	0.003	0.072	0.000	0.003	0.004	0.044	2.608	0.002	0.001	
sim	0.050	3.500	0.025	2.530	-0.650	0.500	30.000	-0.020	0.050	1.000
est	0.051	4.440	0.023	2.415	-0.676	0.417	44.148	-0.024	0.044	
RMSE	0.029	1.049	0.003	0.218	0.057	0.382	35.525	0.020	0.014	
std err	0.003	0.105	0.000	0.022	0.006	0.038	3.552	0.002	0.001	
sim	0.050	3.500	0.025	1.006	-0.650	0.500	30.000	-0.020	0.050	0.750
est	0.050	3.862	0.024	1.052	-0.672	0.509	37.628	-0.023	0.044	0.761
RMSE	0.027	0.840	0.004	0.325	0.050	0.470	29.295	0.020	0.013	0.072
std err	0.003	0.084	0.000	0.032	0.005	0.047	2.929	0.002	0.001	0.007
sim	0.050	3.500	0.025	3.658	-0.650	0.500	30.000	-0.020	0.050	1.100
est	0.052	4.853	0.022	3.334	-0.680	0.424	42.383	-0.023	0.044	1.073
RMSE	0.032	1.343	0.003	1.232	0.072	0.429	32.784	0.025	0.015	0.091
std err	0.003	0.134	0.000	0.123	0.007	0.043	3.278	0.002	0.001	0.009

Table 20: VaR Specification Tests 1% (Full sample)

This table shows model confidence set results for the full out-of-sample period January 3, 2007 to December 30, 2016 using the asymmetric VaR loss function proposed by González-Rivera *et al.* (2004). For details of notation and calculation see Section 4.2 and Section 4.3. A subset of the most relevant models representing each of the model classes analyzed in this paper are tested against each other. The first column indicates the number of the iterative elimination step for models running from $i = 1$ to total number of models ($m_0 = 18$). Second column shows the p-values for the hypotheses $H_{0,\mathcal{M}_i}$ and third column presents MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$ for the model that is going to be eliminated in the respective elimination step. Fourth column shows the model eliminated in each iterative step by the elimination rule and thereby presents the model ranking according the MCS criteria, with the worst model ranked at the top and the best model at the bottom of the table, respectively. For a given significance level α any model for which holds $\hat{p}_{e_{\mathcal{M}_i}} \geq \alpha$ is included in the MCS $\widehat{\mathcal{M}}^*_{1-\alpha}$.

Elimination Rule	p-Value for $H_{0,\mathcal{M}_i}$	MCS p-Value $\hat{p}_{e_{\mathcal{M}_i}}$	Eliminated Model
$e_{\mathcal{M}_1}$	0.1570	0.1570	GJR-MF-J
$e_{\mathcal{M}_2}$	0.2064	0.2064	MF-SV-A
$e_{\mathcal{M}_3}$	0.2159	0.2159	MF-SVJ-A
$e_{\mathcal{M}_4}$	0.2019	0.2159	SVYYD-A
$e_{\mathcal{M}_5}$	0.1882	0.2159	MF-SVJ-G
$e_{\mathcal{M}_6}$	0.1744	0.2159	MF-SVSJJ-A
$e_{\mathcal{M}_7}$	0.1408	0.2159	GJR-MF-N
$e_{\mathcal{M}_8}$	0.1493	0.2159	SVYYD-G
$e_{\mathcal{M}_9}$	0.1868	0.2159	SVSJ-G
$e_{\mathcal{M}_{10}}$	0.2687	0.2687	MF-SV-G
$e_{\mathcal{M}_{11}}$	0.2720	0.2720	SVSJ-A
$e_{\mathcal{M}_{12}}$	0.2990	0.2990	MF-SVSJJ-G
$e_{\mathcal{M}_{13}}$	0.3217	0.3217	GJR-N
$e_{\mathcal{M}_{14}}$	0.4385	0.4385	SV-A
$e_{\mathcal{M}_{15}}$	0.2878	0.4385	SV-G
$e_{\mathcal{M}_{16}}$	0.3565	0.4385	GJR-N-J
$e_{\mathcal{M}_{17}}$	0.4286	0.4385	GJR-MF-t
$e_{\mathcal{M}_{18}}$	1.0000	1.0000	GJR-t